UNIVERSITY OF CENTRAL MISSOURI

Department of Computer Science & Cybersecurity

CHPATER 9. Natural Language Processing

CS 5720: Neural Network & Deep Learning

Dr. I Hua Tsai, Assistant Professor

Welcome to NLP Course! (by ChatGPT!)

Welcome to the Natural Language Processing course!

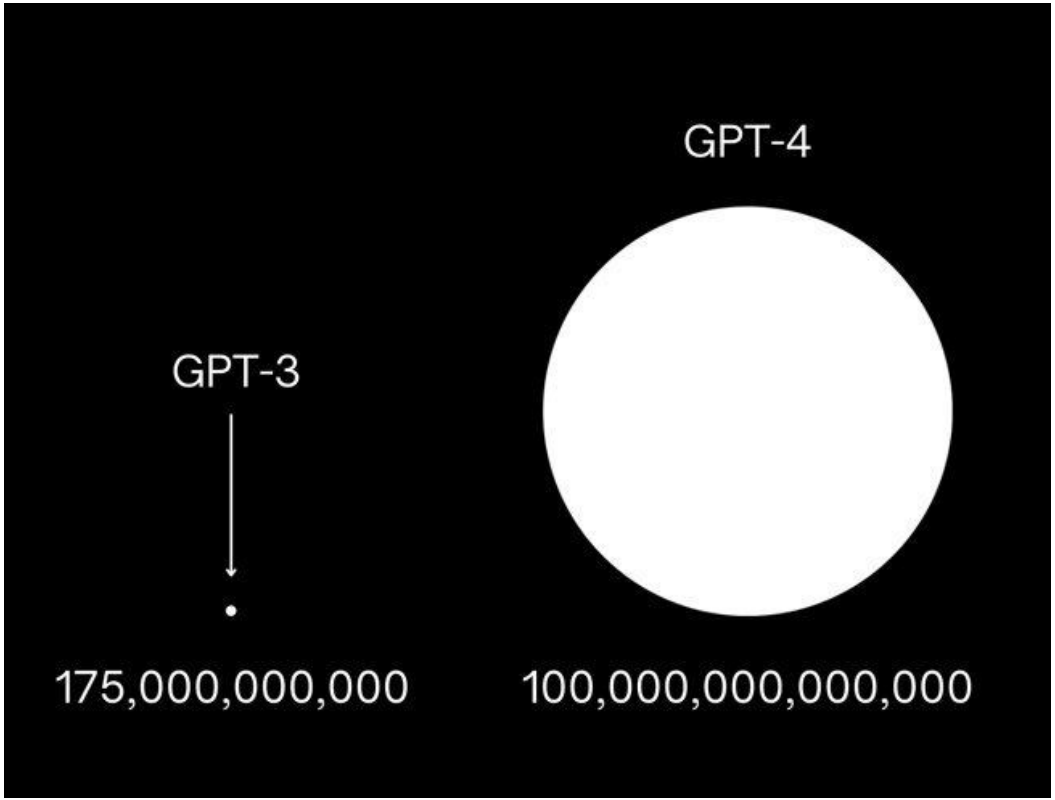
NLP is an exciting and rapidly growing field that deals with the interaction between computers and human language.

In this course, you will learn about the techniques and algorithms used to analyze and understand human language, and you will have the opportunity to apply these techniques to real-world problems.

Whether you are a computer science student, a linguist, or just someone with an interest in language and technology, this course will provide you with a solid foundation in NLP and its applications.

Let's dive in and discover the amazing possibilities of NLP together!

GPT Model	Size (Parameters)	Release Date	Applications
GPT	1.5 billion	June 2017	Text generation, language translation, language modeling, text summarization
GPT-2	1.5 billion	February 2019	Text generation, language translation, language modeling, text summarization
GPT-3	175 billion	June 2020	Text generation, language translation, language modeling, text summarization, question answering, chatbots, automated content generation
CHAT-GPT	175 billion	November 2022	Chatbots, conversation generation
GPT-4	175 billion	Not publicly released	Text generation, language translation, language modeling, text summarization, question answering, chatbots, automated content generation, customer service, education



What is Natural Language Processing (NLP)?

What is Natural Language Processing (NLP)?

Natural language processing is the set of methods for making human language accessible to computers

(Jacob Eisenstein)

Natural language processing is the field at the intersection of Computer science (Artificial intelligence) and linguistics

(Christopher Manning)

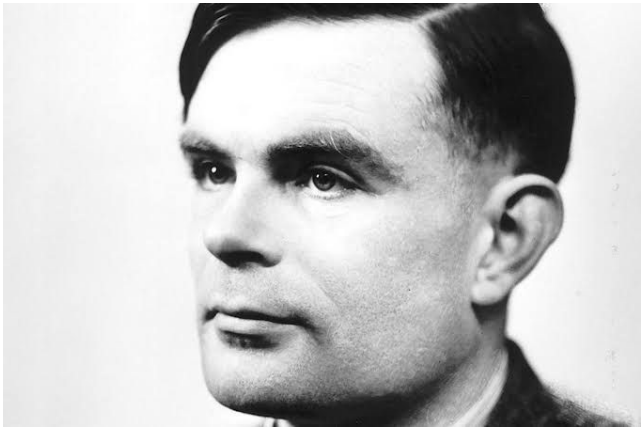
Make computers to understand natural language to do certain task humans can do such as Machine translation, Summarization, Questions answering

(Behrooz Mansouri)

History of NLP

Turing Test

"Computing Machinery and Intelligence"
Mind, Vol. 59, No. 236, pp. 433-460, 1950

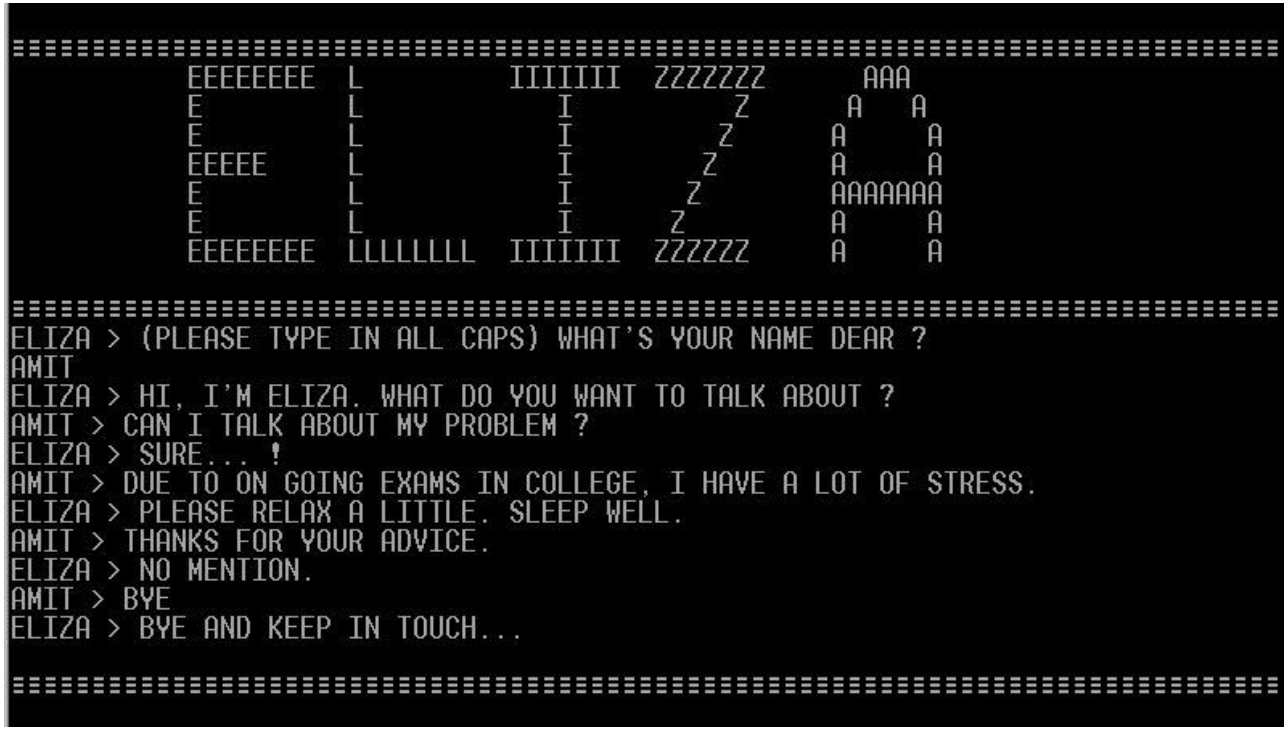


I propose to consider the question
"Can machines think?"...
We can only see a short distance ahead, but
we can see plenty there that needs to be done

In Turing's game, there are three participants: two people and a computer.
One of the people is a contestant who plays the role of an interrogator. To win, the interrogator must determine which of the other two participants is the machine by asking a series of questions via a teletype. The task of the machine is to fool the interrogator into believing it is a person by responding as a person would to the interrogator's questions. The task of the second human participant is to convince the interrogator that the other participant is the machine and that she is human.

Q: Please write me a sonnet on the topic of the Forth Bridge.
A: Count me out on this one. I never could write poetry.
Q: Add 34957 to 70764.
A: (Pause about 30 seconds and then give answer as) 105621.

ELIZA



ELIZA was an early natural language processing system capable of carrying on a limited form of conversation with a user

1950 – 1970

Mid 1950's – Mid 1960's: Birth of NLP and Linguistics

- At first, people thought NLP is easy! Researchers predicted that "machine translation" can be solved in 3 years or so
- Mostly hand-coded rules / linguistic-oriented approaches
- The 3-year project continued for 10 years, but still no good result, despite the significant amount of expenditure

Mid 1960's – Mid 1970's: A Dark Era

- After the initial hype, a dark era follows
- People started believing that machine translation is impossible, and most abandoned research for NLP

1970 – 2000

1970's and early 1980's – Slow Revival of NLP

- □Some research activities revived, but the emphasis is still on linguistically oriented, working on small toy problems with weak empirical evaluation

Late 1980's and 1990's – Statistical Revolution!

- □By this time, the computing power increased substantially
- □Data--driven, statistical approaches with simple representation win over complex hand-coded linguistic rules
- "Whenever I fire a linguist, our machine translation performance improves." (Jelinek, 1988)

2000's – Statistics Powered by Linguistic Insights

- □With more sophistication with the statistical models, richer linguistic representation starts finding a new value

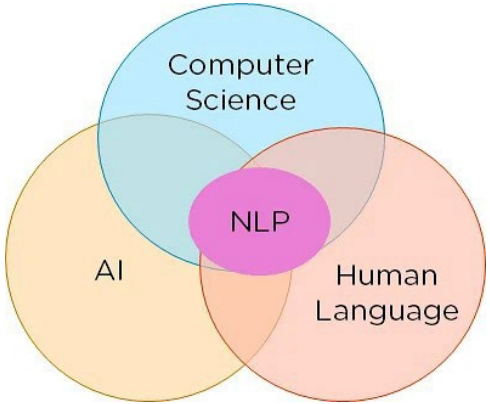
Recent Years

2010's – Emergence of embedding model and deep neural networks

- □Several embedding models for text using neural networks and deep neural networks were proposed including Word2Vec, Glove, fastText, Elmo, BERT, COLBERT, GTP[1-3.5]
- New techniques brought attention to more complex tasks

Natural Language Processing

NLP combines the field of linguistics and computer science to decipher language structure and guidelines and to make models which can comprehend, break down and separate significant details from text and speech.

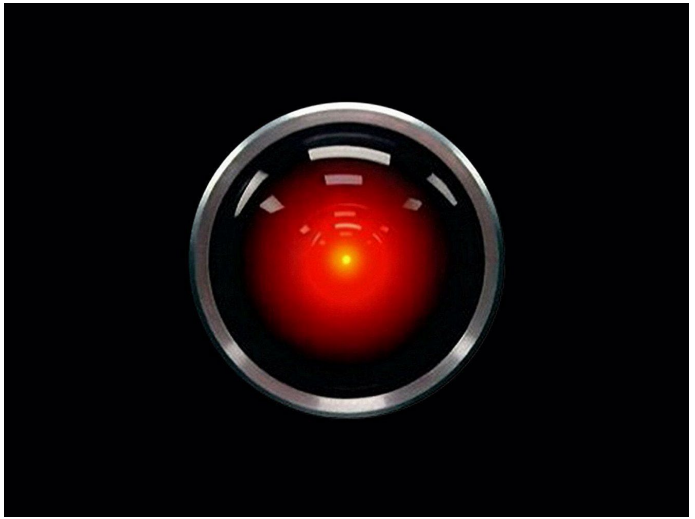


Example: Conversational Agent

Conversational agents contain:

- Speech recognition
- Language analysis
- Dialogue processing
- Information retrieval
- Text to speech

David Bowman:
Open the pod bay doors, Hal.
HAL:
I'm sorry, Dave, I'm afraid I can't do that.
David Bowman:
What are you talking about, Hal?
...HAL:
I know that you and Frank were planning to disconnect me, and I'm afraid that's something I cannot allow to happen.



2001: A Space Odyssey – HAL 9000

HAL is an artificial agent capable of such advanced language-processing behavior as speaking and understanding English, and at a crucial moment in the plot, even reading lips

Natural Language Processing: Terms

Natural language refers to the language that humans use to communicate with each other, such as English, Spanish, or Chinese

Processing

As distinguished from data processing

Question: How is data processing and natural language processing different?

Natural Language Processing: Terms

Consider the Unix wc program, which counts the total number of bytes, words, and lines in a text file

- When used to count bytes and lines, wc is an ordinary data processing application
- However, when it is used to count the words in a file, it requires knowledge about what it means to be a word and thus becomes a language processing system

Natural Language Processing vs Computational Linguistics

In linguistics, language is the object of study

- Computational methods may be brought to bear, just as in scientific disciplines like computational biology and computational astronomy, but they play only a supporting role

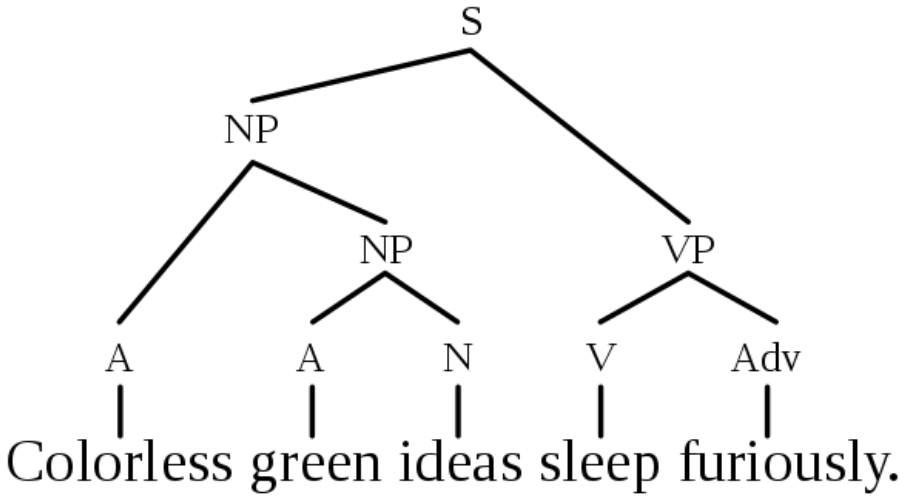
In contrast, natural language processing is focused on the design and analysis of computational algorithms and representations for processing natural human language

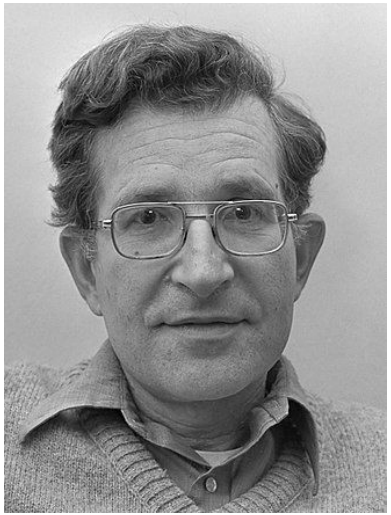
- The goal of natural language processing is to provide new computational capabilities around human language: for example, extracting information from texts, translating between languages, answering questions, holding a conversation, taking instructions

Syntax vs. Semantics

Colorless green ideas sleep furiously.

(example by Noam Chomsky 1957)





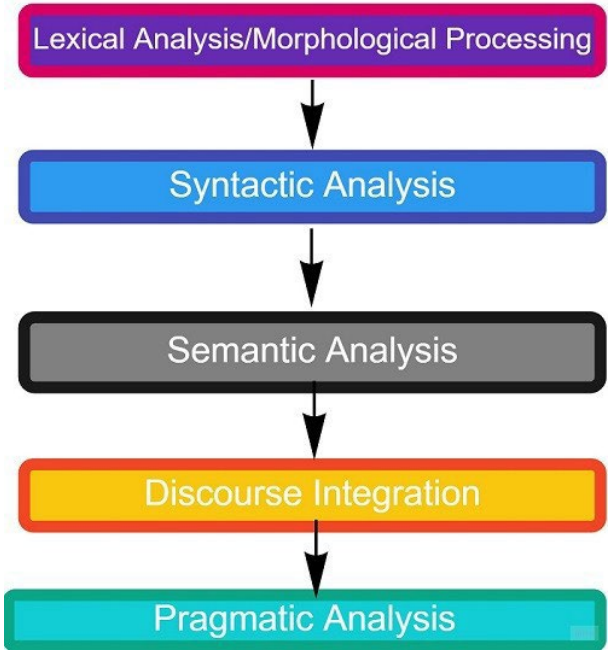
Noam Chomsky
The most cited person alive

Semantics vs. Pragmatics

What does "You have a green light" mean?

- □ You are holding a green light bulb?
- □ You have a green light to cross the street?
- □ You can go ahead with your plan?

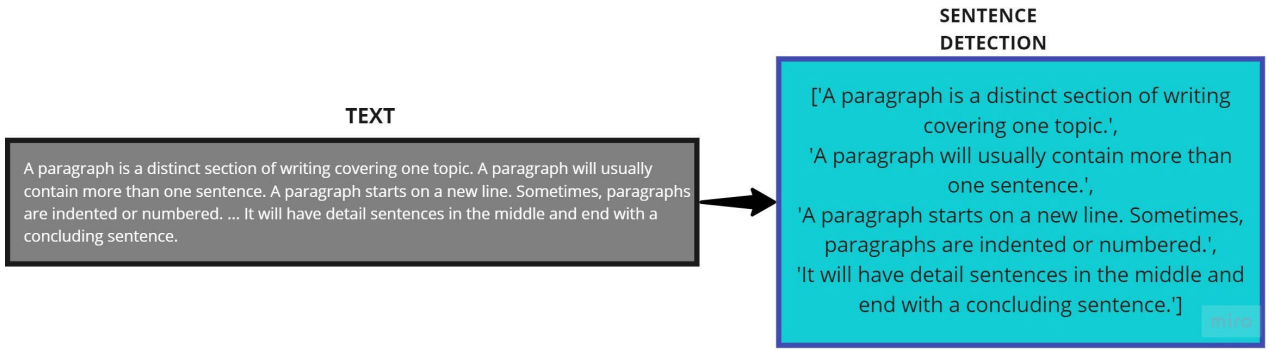
Phases of NLP



Lexical Analysis

- The first phase is lexical analysis/morphological processing. In this phase, the sentences, paragraphs are broken into tokens.
- These tokens are the smallest unit of text. It scans the entire source text and divides it into meaningful lexemes.
- For example, The sentence “He goes to college.” is divided into [‘He’ , ‘goes’ , ‘to’ , ‘college’, ‘.’] .
- There are five tokens in the sentence. A paragraph may also be divided into sentences.

Lexical Analysis



Syntactic Analysis/Parsing

- The second phase is Syntactic analysis. In this phase, the sentence is checked whether it is well- formed or not.
- The word arrangement is studied and a syntactic relationship is found between them. It is checked for word arrangements and grammar.
- For example, the sentence “Delhi goes to him” is rejected by the syntactic parser.

Semantic Analysis

- The third phase is Semantic Analysis. In this phase, the sentence is checked for the literal meaning of each word and their arrangement together.
- For example, The sentence “I ate hot ice cream” will get rejected by the semantic analyzer because it doesn’t make sense.
- E.g.. “colorless green idea.” This would be rejected by the Symantec analysis as colorless Here; green doesn’t make any sense.

Discourse Integration

- The fourth phase is discourse integration. In this phase, the impact of the sentences before a particular sentence and the effect of the current sentence on the upcoming sentences is determined.
- For example, the word “that” in the sentence “He wanted that” depends upon the prior discourse context.

Pragmatic Analysis

- The last phase of natural language processing is Pragmatic analysis. Sometimes the discourse integration phase and pragmatic analysis phase are combined.
- The actual effect of the text is discovered by applying the set of rules that characterize cooperative dialogues.
- E.g., “close the window?” should be interpreted as a request instead of an order.

NLP Implementation

Below, given are popular methods used for Natural Learning Process:

- Machine learning: The learning nlp procedures used during machine learning. It automatically focuses on the most common cases. So when we write rules by hand, it is often not correct at all concerned about human errors.
- Statistical inference: NLP can make use of statistical inference algorithms. It helps you to produce models that are robust. e.g., containing words or structures which are known to everyone.

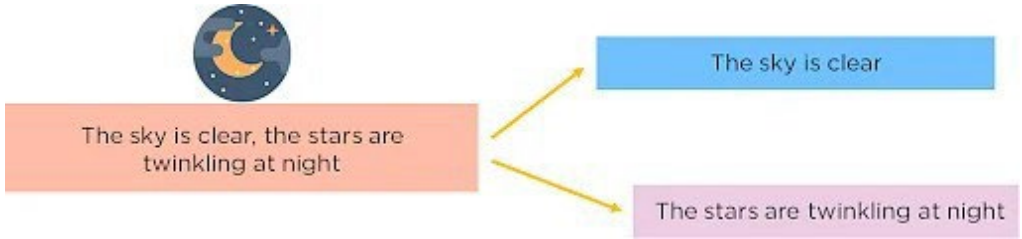
NLP Steps

How to Perform NLP?

- Segmentation
- Tokenizing
- Removing Stop Words:
- Stemming
- Lemmatization
- Part of Speech Tagging
- Named Entity Tagging

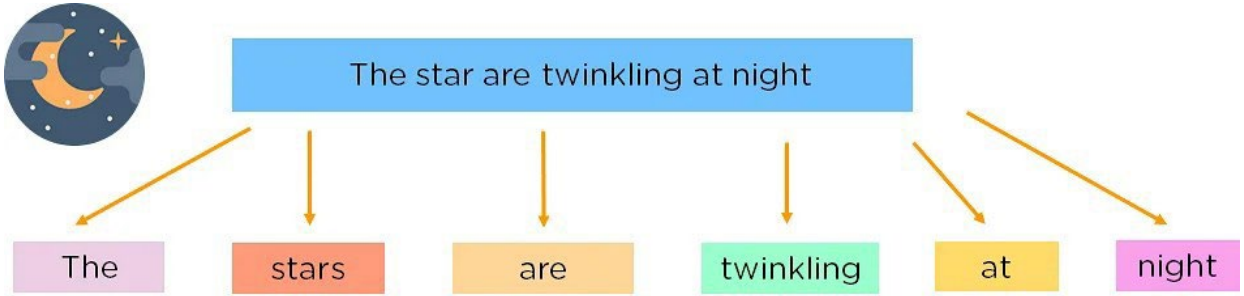
Segmentation

- You first need to break the entire document down into its constituent sentences. You can do this by segmenting the article along with its punctuation like full stops and commas.



Tokenizing

- For the algorithm to understand these sentences, you need to get the words in a sentence and explain them individually to our algorithm.
- So, you break down your sentence into its constituent words and store them. This is called tokenizing, and each world is called a token.



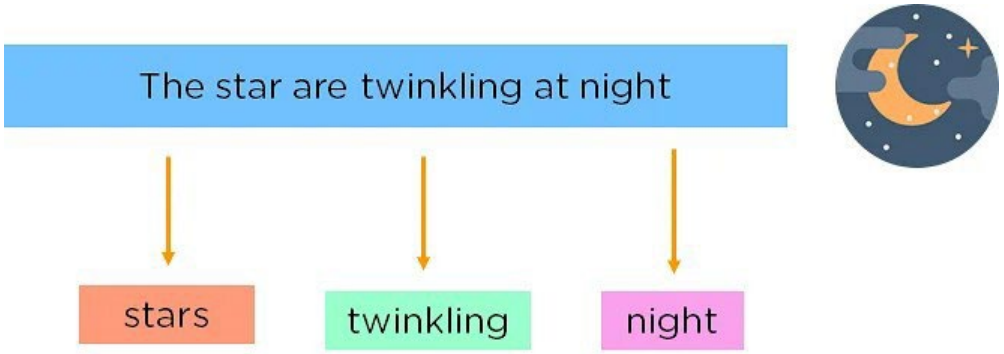
Example

```
from nltk.tokenize import regexp_tokenize

sentence = " students learn NLP concepts quickly!"
tokens = regexp_tokenize(sentence, pattern=r"\w+")
print(tokens)
```

```
['students', 'learn', 'NLP', 'concepts', 'quickly']
```

Removing Stop Words

- You can make the learning process faster by getting rid of non-essential words, which add little meaning to our statement and are just there to make our statement sound more cohesive. Words such as was, in, is, and, the, are called stop words and can be removed.



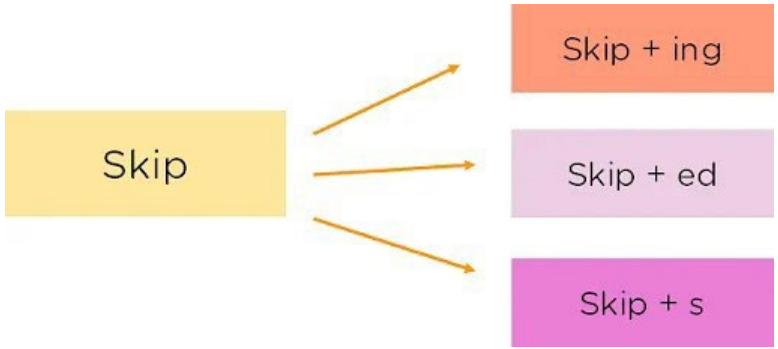
Example

```
import nltk
#nltk.download('all')  # to download everything you'll need for the course


from nltk.corpus import stopwords
stop_words = set(stopwords.words('english'))

# Example usage
from nltk.tokenize import word_tokenize
nltk.download('punkt')  # just in case it wasn't done yet

sentence = " students is the learn NLP concepts quickly!"
tokens = word_tokenize(sentence)

filtered_tokens = [word for word in tokens if word.lower() not in stop_words]
print("Filtered tokens:", filtered_tokens)
```

```
Filtered tokens: ['students', 'learn', 'learn', 'NLP', 'concepts', 'quickly', '!']
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Stemming

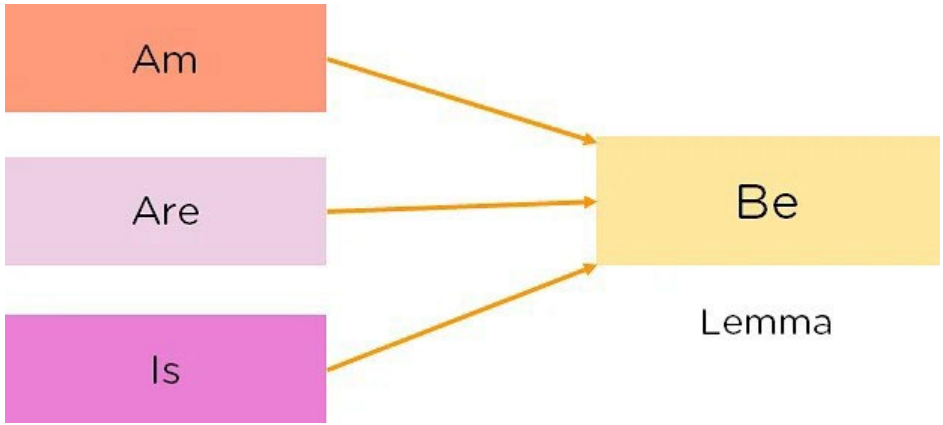
- It is the process of obtaining the Word Stem of a word. Word Stem gives new words upon adding affixes to them



Lemmatization

- The process of obtaining the Root Stem of a word. Root Stem gives the new base form of a word that is present in the dictionary and from which the word is derived.

You can also identify the base words for different words based on the tense, mood, gender,etc.



Feature	Stemming	Lemmatization
Rule-based	Heuristics	Dictionary + Grammar
Output	Might be invalid	Always valid word
Speed	Faster	Slower
Accuracy	Lower	Higher

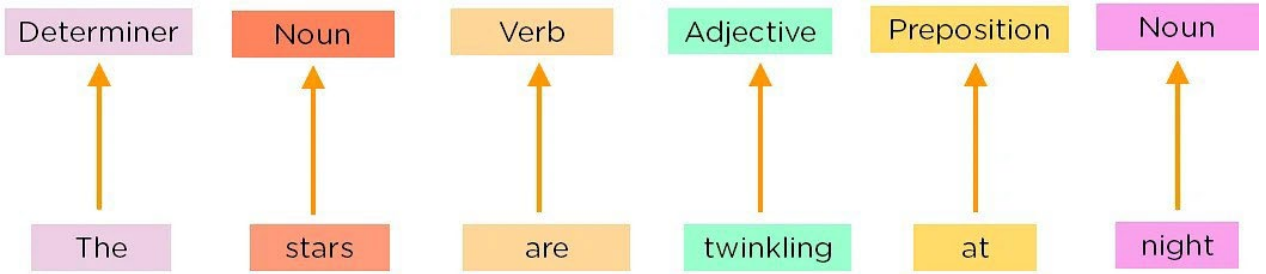
Example

```
from nltk.stem import PorterStemmer, WordNetLemmatizer
stemmer = PorterStemmer()
lemmatizer = WordNetLemmatizer()

words = ["running", "flies", "better", "happily"]

print("Stemming:")
print([stemmer.stem(word) for word in words])

print("Lemmatization:")
print([lemmatizer.lemmatize(word, pos='v') for word in words])
```

```
Stemming:
['run', 'fli', 'better', 'happili']
Lemmatization:
['run', 'fly', 'better', 'happily']
```

Part of Speech Tagging

- Now, you must explain the concept of nouns, verbs, articles, and other parts of speech to the machine by adding these tags to our words. This is called ‘part of’.



Example

```
from nltk import pos_tag
print(pos_tag(tokens))
```

```
[('students', 'NNS'), ('learn', 'VBP'), ('NLP', 'NNP'), ('concepts', 'NNS'), ('quickly', 'RB'), ('!', '.')]
```

Named Entity Tagging

- Next, introduce your machine to pop culture references and everyday names by flagging names of movies, important personalities or locations, etc that may occur in the document.
- You do this by classifying the words into subcategories. This helps you find any keywords in a sentence. The subcategories are person, location, monetary value, quantity, organization, movie.
- After performing the preprocessing steps, you then give your resultant data to a machine learning algorithm like Naive Bayes, etc., to create your NLP application.

Example

```
import spacy
nlp = spacy.load("en_core_web_sm")
doc = nlp("Barack Obama was the 44th President of the United States.")

for ent in doc.ents:
    print(ent.text, ent.label_)
```

```
Barack Obama PERSON
44th ORDINAL
the United States GPE
```

In-class assignment

Is NLP hard?

What does this sentence mean? “*I made her duck*”

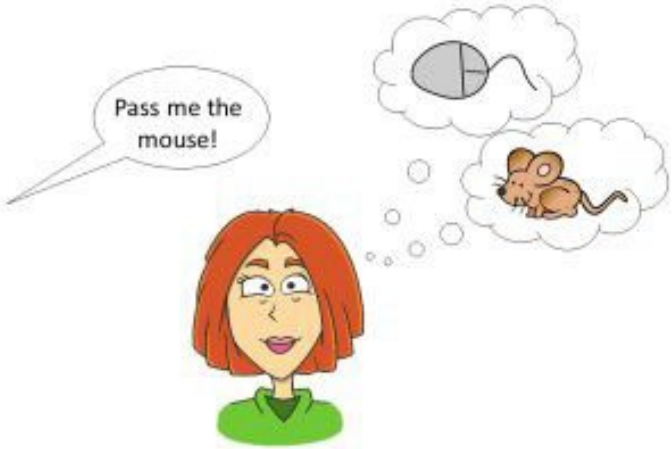
“duck”: noun or verb?
“make”: “cook X” or “cause X to do Y” ?
“her”: “for her” or “belonging to her” ?

- I cooked waterfowl for her
- I cooked waterfowl belonging to her
- I created the (plaster?) duck she owns
- I caused her to quickly lower her head or body
- I waved my magic wand and turned her into undifferentiated waterfowl

These different meanings are caused by a number of ambiguities

- First, the words duck and her are morphologically or syntactically ambiguous in their part-of-speech
  - Duck can be a verb or a noun, while her can be a dative pronoun or a possessive pronoun
- Second, the word make is semantically ambiguous; it can mean create or cook
- Finally, the verb make is syntactically ambiguous in a different way

We Need to Disambiguate



Disambiguation

Models and algorithms in this course are ways to resolve or disambiguate these ambiguities

- Deciding whether duck is a verb or a noun can be solved by part-of-speech tagging
- Deciding whether make means “create” or “cook” can be solved by word sense disambiguation

Resolution of part-of-speech and word sense ambiguities are two important kinds of lexical disambiguation

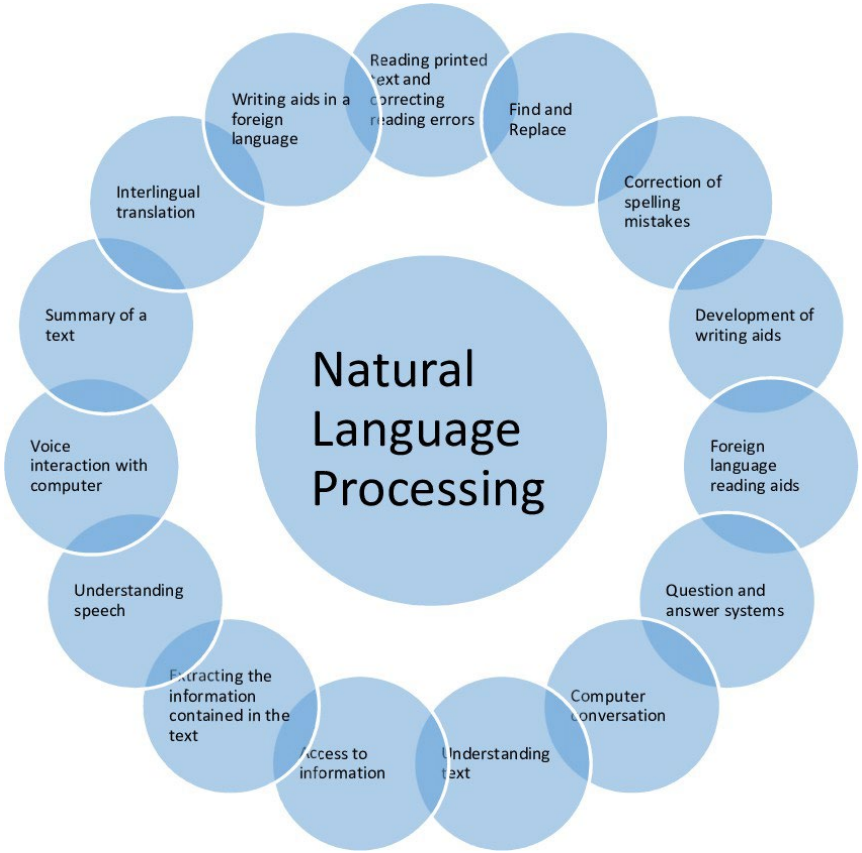
A wide variety of tasks can be framed as lexical disambiguation problems

- A text-to-speech synthesis system reading the word lead needs to decide whether it should be pronounced as in lead pipe or as in lead me on
- Deciding whether her and duck are part of the same entity or are different entities is an example of syntactic disambiguation and can be addressed by probabilistic parsing

Tasks/Applications in NLP

A few of the NLP Tasks

- Spell Checking, Keyword Search, Finding Synonyms
- Part of Speech Tagging
- Extracting information from a website
  - Location, people, temporal expressions
- Classifying text
  - Sentiment analysis
- Machine translation
- Complex question answering
- Spoken dialog systems


Natural Language Processing
Reading printed text and correcting reading errors
Find and Replace
Correction of spelling mistakes
Development of writing aids
Foreign language reading aids
Question and answer systems
Computer conversation
Understanding text
Access to information
Extracting the information contained in the text
Understanding speech
Voice interaction with computer
Summary of a text
Interlingual translation
Writing aids in a foreign language

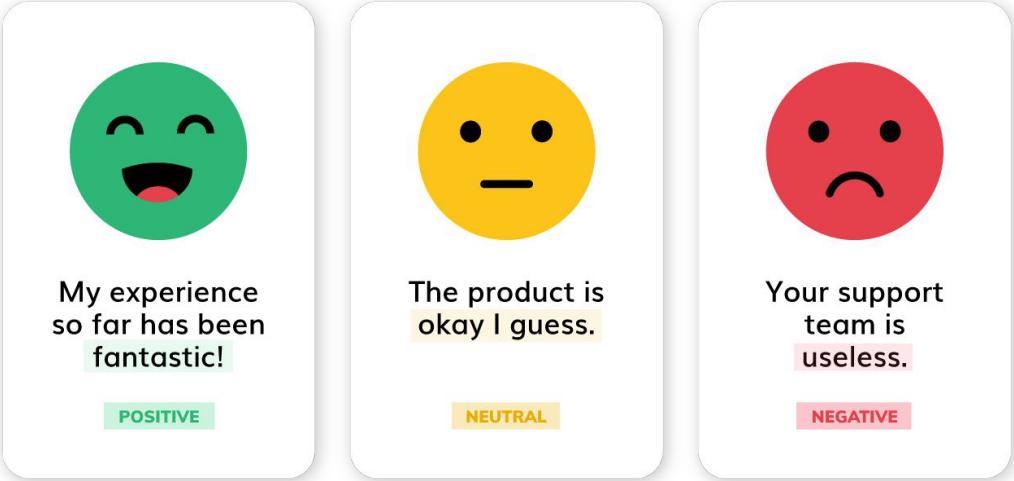
Knowledge & Information Extraction

Knowledge graphs (KGs) organize data from multiple sources, capture information about entities of interest in a given domain or task (like people, places or events), and forge connections between them

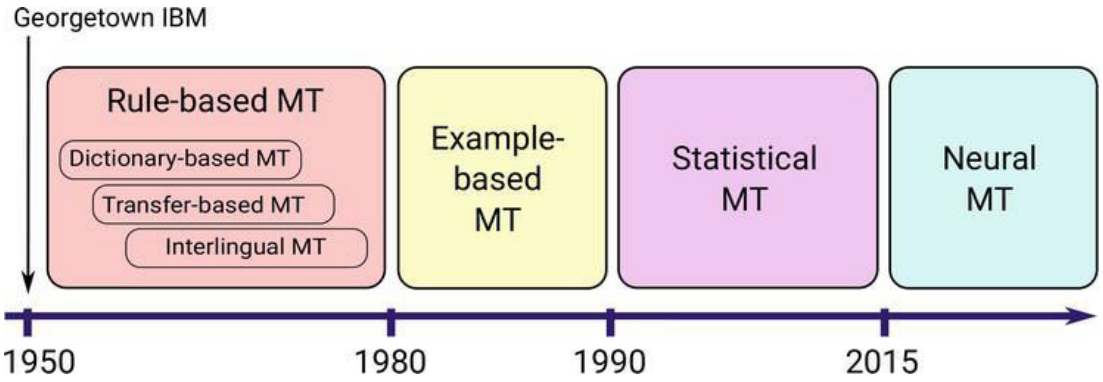
The Google Knowledge Graph is an enormous database of information that enables Google to provide immediate, factual answers to your questions

Sentiment Analysis

Determine whether the meaning behind data is positive, negative, or neutral



Machine Translation



Low resource languages can be challenging?

6,800 living languages
600 with written tradition
100 spoken by 95% of population

Question Answering

IBM-Watson Defeats Humans in "Jeopardy!"

Spoken Dialog Systems

How ChatGPT Works in NLP

1.Pre-training:

1. The model is trained on a large dataset containing diverse text sources (books, articles, conversations).
2. It learns grammar, facts, reasoning, and linguistic patterns.

2.Fine-tuning:

1. Further refined using **supervised learning** and **reinforcement learning from human feedback (RLHF)**.
2. This process helps align responses with user expectations, making them more useful and ethical.

3.Inference (Generation):

1. Given a prompt, the model predicts the most likely next words based on learned patterns.
2. It uses **context and attention mechanisms** to maintain coherence and logical flow.

Algorithm / Model	NLP Task	Description	Typical Use Case
Hidden Markov Model (HMM)	POS Tagging / Sequence Labeling	Probabilistic model for sequence prediction	Speech tagging, entity recognition
Naive Bayes Classifier	Text Classification / Sentiment	Probabilistic classifier based on Bayes theorem with strong independence assumption	Spam detection, sentiment analysis
BERT (Transformer-based)	Multiple NLP Tasks (QA, NER, etc.)	Contextual embeddings using bidirectional attention	Question answering, sentiment analysis, NER
GPT (Generative Pretrained Transformer)	Language Modeling / Generation	Transformer model trained for text generation and understanding	Text generation, summarization, chatbots
SpaCy / NLTK POS Taggers	Part-of-Speech Tagging	Rule-based and statistical taggers	Grammar analysis, text annotation

Advantage

Advantage	Description
Enhanced Data Processing	Automates the extraction and processing of large volumes of text efficiently.
Improved User Experience	Powers chatbots and virtual assistants, making interactions more natural.
Insight Extraction	Enables sentiment analysis, topic modeling, and other forms of deep insight.
Multilingual Support	Facilitates translation and cross-language communication.
Automation	Automates routine tasks like summarization, classification, and information retrieval.

Disadvantage

Disadvantage	Description
Ambiguity	Natural language is inherently ambiguous, which can lead to misinterpretations.
Data Bias	Models may inherit and even amplify biases present in the training data.
Complexity	Requires extensive data and computational resources for training robust models.
Context Limitations	Struggles with nuanced contexts, cultural subtleties, and long-term dependencies.
Error Propagation	Mistakes in preprocessing or interpretation can lead to compounding errors.

Future

Future Trend	Description
Model Efficiency	Focus on creating lighter, faster models that can run in real time.
Multimodal Learning	Integrating text with images, audio, and video for a richer contextual understanding.
Explainable AI	Enhancing transparency in model decisions to make them more interpretable.
Continual Learning	Developing systems that adapt continuously to new data and evolving language.
Ethical and Fair AI	Prioritizing methods that reduce bias and ensure ethical application of NLP.