Deep Dive into *Natural Language Processing*

Presented by Tanya Khanna

WORKSHOPS SCHEDULE

https://libcal.rutgers.edu/calendar/nblworkshops?cid=4537&t=d&d=0000-00-00&cal=4537&inc=0

https://github.com/Tanya-Khanna/DataScienceWorkshop_2024_NBL

Introduction to Python Programming	February 1, 2024; 4 - 5:30 PM
Mastering Data Analysis: Pandas & Numpy	February 8, 2024; 4 - 5:30 PM
Python for Visualization & Exploration	February 15, 2024; 4 - 5:30 PM
Mathematical Foundations of Data Science	February 29, 2024; 4 - 5:30 PM
Introduction to Machine Learning: Supervised	March 7, 2024; 4 - 5:30 PM
Introduction to Machine Learning: Unsupervised	March 21, 2024; 4 - 5:30 PM
Introduction to Deep Learning	March 28, 2024; 4 - 5:30 PM
Deep Dive into Natural Language Processing	April 4, 2024; 4 - 5:30 PM
Large Language Models and Chat GPT	April 11, 2024; 4 - 5:30 PM

Find something which is *common* for all below?

- Wharton MBA exam
- Law School Exam
- Stanford Medical School clinical reasoning final
- US medicine licensing exam

American Medical Association
https://www.ama-assn.org › ... › Digital

ChatGPT passed the USMLE. What does it mean for med ed?

Mar 3, 2023 — A recently published study has spotlighted its ability to pass well-known **licensing exams**, suggesting a useful role in **medical** education.

Knowledge at Wharton
https://knowledge.wharton.upenn.edu › podcast › chatg...

ChatGPT Passed an MBA Exam. What's Next?

Jan 31, 2023 — Terwiesch's white paper has garnered media attention with its intriguing title, "Would **Chat GPT** Get a **Wharton MBA**?" The answer is a solid "yes," ...

CNN
https://www.cnn.com › 2023/01/26 › tech › chatgpt-pas...

ChatGPT passes exams from law and business schools

Jan 26, 2023 — The powerful new AI chatbot tool recently passed **law exams** in four courses at the **University** of Minnesota and another **exam** at **University** of ...

Stanford HAI
https://hai.stanford.edu › news › chatgpt-out-scores-me...

ChatGPT Out-scores Medical Students on Complex Clinical ...

Jul 17, 2023 — **ChatGPT** can outperform first- and second-year **medical students** in answering challenging clinical care **exam** questions, a new study by **Stanford** ...

TABLE OF CONTENTS

NLP: Bridging Humans and Machines

Core Components of NLP

Major NLP Techniques

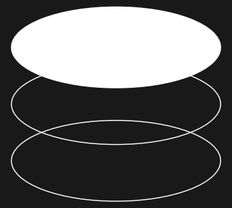
Data Preprocessing: Preparing Text for Insight

Sentiment Analysis: How machines understand the subtleties of human emotion through text

Topic Modelling: Techniques for discovering the hidden themes within large volumes of text

Text Generation: A look at how NLP models create text that mimics human writing styles and thought processes

What is Natural Language Processing (NLP)?



NLP is a branch of artificial intelligence (AI) that gives computers the ability to understand, interpret, and produce human languages. The main goal of NLP is to bridge the gap between human communication and computer understanding.

Why is NLP important?

2.5

quintillion bytes of data generated daily - a significant portion of which is unstructured text, NLP allows for the automated analysis of this vast volume of text data. This capability aids in market research, customer service, and more by extracting valuable insights from online reviews, social media posts, customer feedback, etc.

85%

of smartphone users utilize voice assistants, like Siri and Alexa, for various tasks. Predictive text, powered by NLP, assists in over 60% of mobile communications, enhancing typing efficiency and accuracy.

20%

of users are with disabilities, NLP makes voice-controlled assistance technology more accessible, providing a hands-free experience and aiding in navigation, information retrieval, and daily tasks. Also, NLP-powered translation services cover over 100 languages, significantly enhancing accessibility for global communication.

History and Evolution of NLP

ELIZA (1966)

The first rule based chatbot, ELIZA, is created by Joseph Weizenbaum, simulating a psychotherapist.

Rise of Statistical Methods (1980)

Transition to statistical methods in NLP, focusing on machine learning algorithms for language processing.

WordNet (1995)

A large lexical database of English, marking significant progress in linguistic databases.

Stanford Parser (2003)

An important tool for linguistic analysis and parsing sentences.

Seq2Seq Models (2014)

Introduction of sequence-to-sequence models, enhancing translation and text summarization.

BERT (2018)

Google introduces BERT, a breakthrough in machine understanding of context in text.

GPT-3 (2020)

OpenAI releases GPT-3, a state-of-the-art language model with broad applications in writing, conversation, and more.

Large-scale Multimodal Models (2021-2023)

Advancements in integrating text with other data types (e.g., images, videos) to understand and generate multimodal content.

Core Components of NLP

Text Parsing

It is the process of analyzing a text's structure and recognizing its meaningful parts. It involves breaking down text into tokens (such as words and phrases) and identifying its grammatical structure. This is the first step in understanding the text at a deeper level.

Syntactic Analysis (Syntax)

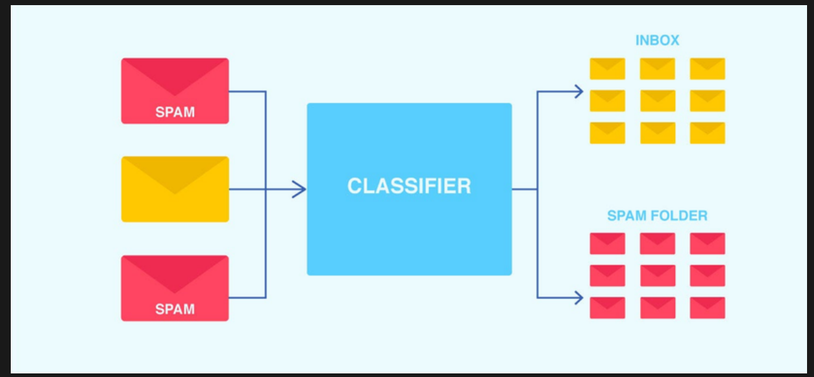
This examines how words are organized into sentences, looking at the grammatical structure. This process identifies the relationships between words, such as subjects, predicates, objects, and other elements of a sentence, to understand the rules that govern sentence construction.

Semantic Analysis (Semantics)

This focuses on the meaning of individual words, phrases, sentences, and the text as a whole. It interprets the meanings that language conveys, beyond just the dictionary definitions of words, considering context and how the meaning of sentences changes with different word arrangements.

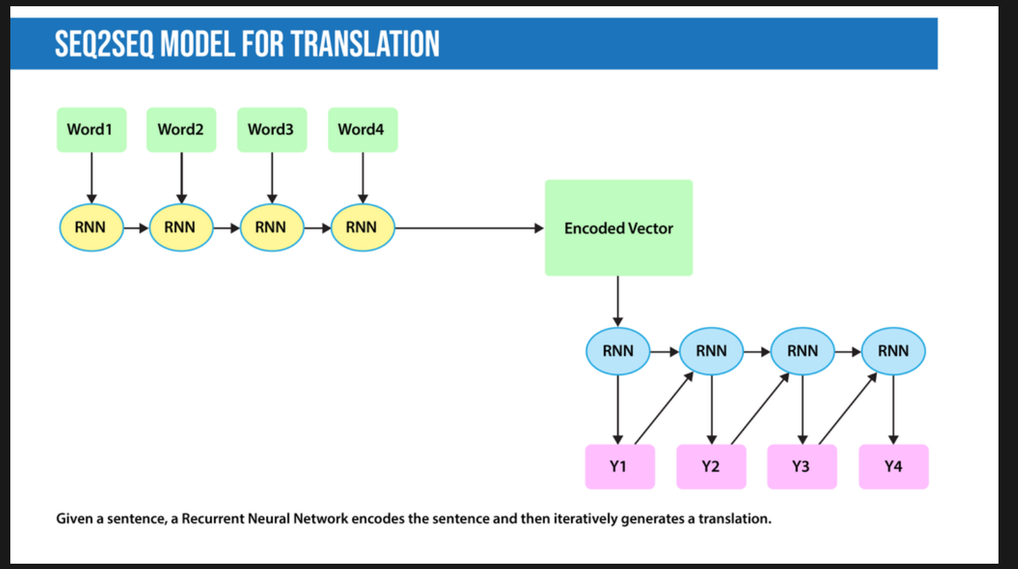
Machine Learning in *NLP*

- Machine learning in NLP involves using algorithms to analyze, understand, and generate human language. The models learn from vast amounts of text data to perform tasks like translation, sentiment analysis, and more.
- **Pre-Deep Learning Era:** Before deep learning took the stage, machine learning in NLP relied on models like decision trees, support vector machines (SVMs), and linear regression, often with handcrafted features.
- **Examples:**
  - Spam Detection: Early machine learning models were trained to identify and filter out spam emails with high accuracy.
  - Part-of-Speech Tagging: Tagging words in a sentence as nouns, verbs, adjectives, etc., using algorithms like Hidden Markov Models (HMMs).
- **Impact:** These models significantly improved the automation of text analysis tasks, reducing the reliance on rule-based systems.

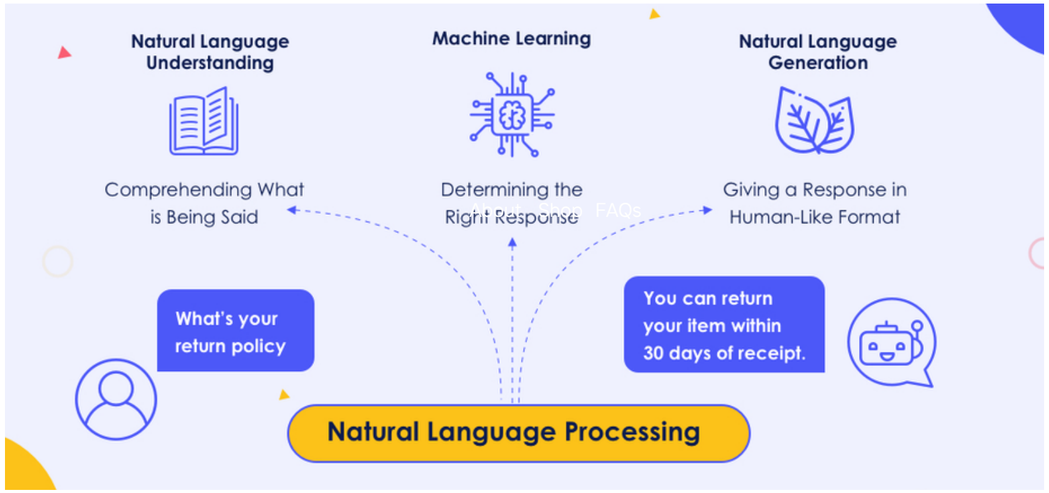


Deep Learning in *NLP*

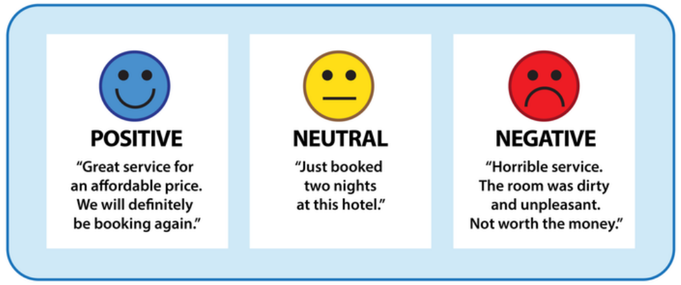
- Deep learning, a subset of machine learning, uses neural networks with many layers (deep architectures) to learn representations of data. In NLP, this has led to models that understand language context and nuances more effectively than ever before.
- Breakthrough Models
  - Recurrent Neural Networks (RNNs): Suitable for sequential data like text, allowing previous outputs to influence the next input.
  - Transformers and Attention Mechanisms: Introduced by the paper "Attention is All You Need" in 2017, transformers revolutionized NLP by enabling models to focus on different parts of the input data, leading to more context-aware processing.
- **Examples:**
  - Machine Translation: Google Neural Machine Translation system uses deep learning for more accurate and contextually relevant translations.
  - Natural Language Understanding: BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) series have set new standards for understanding context and generating human-like text.
- **Impact:** Deep learning has made it possible to process and analyze language in a way that is much closer to human understanding, paving the way for advanced applications like real-time multilingual communication, highly interactive chatbots, and more sophisticated sentiment analysis.



Major *NLP Techniques*



SENTIMENT ANALYSIS



Given text, sentiment analysis classifies its emotional quality.

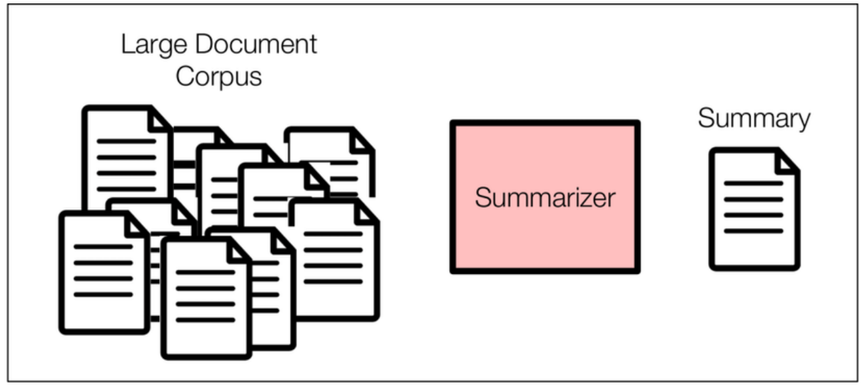
NAMED ENTITY RECOGNITION (NER) TAGGING

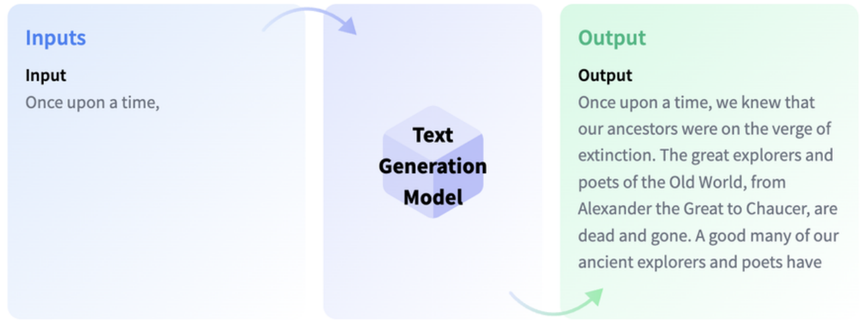
Andrew Yan-Tak Ng PERSON (Chinese NORP : 吳恩達; born 1976 DATE) is a British NORP -born American NORP computer scientist and technology entrepreneur focusing on machine learning and AI GPE . Ng was a co-founder and head of Google Brain ORG and was the former chief scientist at Baidu ORG , building the company's Artificial Intelligence Group ORG into a team of several thousand CARDINAL people.

spaCy named entity recognition tagging of the first paragraph of Andrew Ng's Wikipedia page. "NORP" stands for nationalities or religious or political groups. Note that spaCy incorrectly labels "AI" as "GPE," for geopolitical entity.

NLU

NLG





Applications of NLP

Virtual Assistants

NLP powers virtual assistants like Siri, Alexa, and Google Assistant, allowing users to interact with their devices using natural language.

- Asking Siri to set a reminder
- Requesting Alexa to play a specific song
- Asking Google Assistant for the weather forecast

Sentiment Analysis

NLP can analyze text to determine the sentiment expressed, helping companies understand customer opinions and feedback.

- Analyzing social media posts to gauge customer satisfaction
- Monitoring product reviews to identify positive or negative sentiment
- Assessing customer support chat logs for sentiment analysis

Language Translation

NLP enables automatic language translation, making it easier for people to communicate across different languages.

- Using Google Translate to translate a webpage
- Translating text messages in real-time using language translation apps
- Converting subtitles of a foreign movie to your native language

Text Summarization

NLP can automatically generate summaries of long texts, saving time and providing concise information.

- Using an app to summarize news articles
- Generating abstracts for research papers
- Creating brief descriptions of long emails

Spam Detection

NLP helps identify and filter out spam emails, ensuring that important messages reach the intended recipients.

- Email providers automatically moving suspicious emails to the spam folder
- Flagging potential phishing emails based on their content
- Detecting and blocking spam comments on websites

Challenges in NLP

Ambiguity

Language is inherently ambiguous. Words can have multiple meanings (polysemy), sentences can be interpreted in different ways, and pronouns can have ambiguous antecedents, making it difficult for NLP systems to determine the correct interpretation.

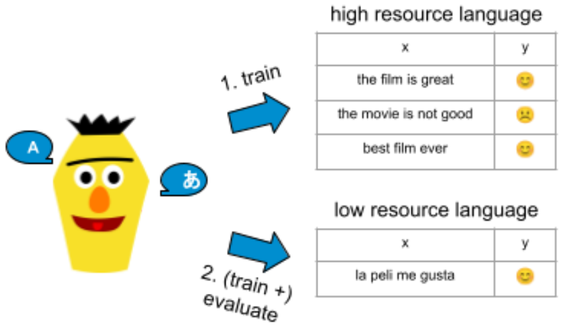
Does a word clever have a negative nuance sometimes? Is there any difference between two sentences below? He is smart. and He is clever.

Understanding Nuances

Capturing the subtleties such as emotion, tone, and nuance in text is a complex task. This includes determining the author's intentions, the strength of their opinions, or the mood they are trying to convey.

Data Bias and Fairness

Machine learning models in NLP can inadvertently learn and perpetuate biases present in their training data. Ensuring that NLP systems are fair and unbiased is a significant challenge.



Lack of Resources for Low-Resource Languages

While NLP has made great strides in languages with abundant data (like English), there is a lack of annotated datasets and resources for many lesser-spoken languages, which hampers the development of NLP technologies for those languages.