Natural Language Processing

WBAI059-05

university of groningen / faculty of science and engineering

Tsegaye Misikir Tashu

Lecture 1: Introduction to Natural Language Processing

Lecture Plan

Lecture 1: Introduction to Natural Language Processing

1. The course Structure
2. What is NLP
3. The challenges
4. Key NLP **use cases and** components
5. Tools for NLP

What do we hope to teach?

1. The foundations of the effective methods of modern NLP
   - Basics first, then key methods used in NLP: Word vectors, feed-forward networks, recurrent neural networks, encoder-decoder models, attention, transformers

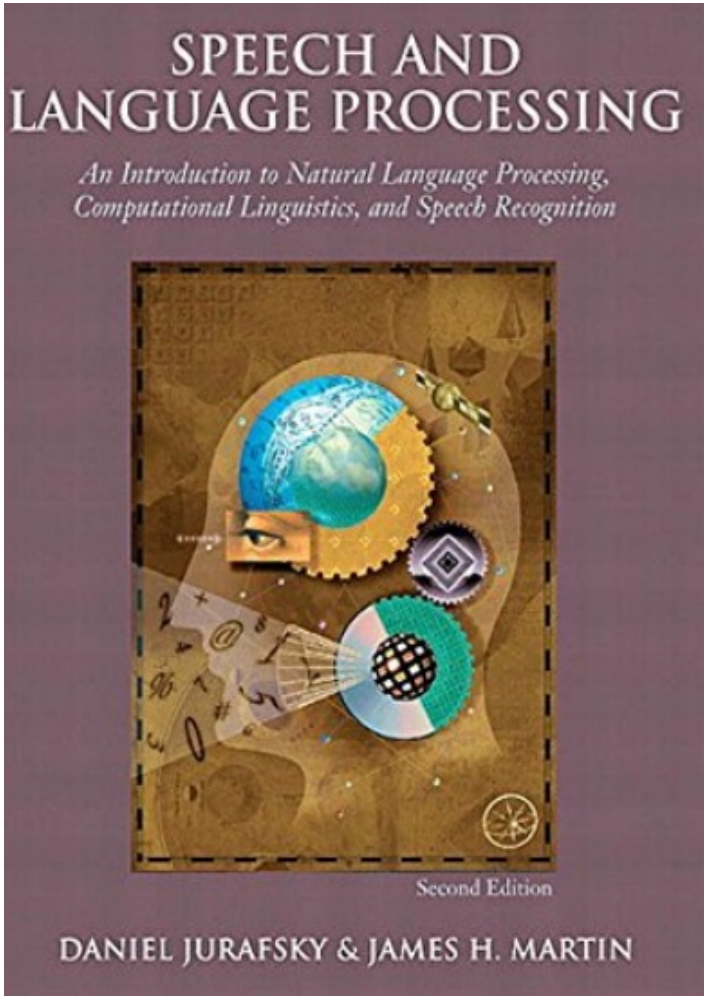
2. An understanding of and **ability to build systems** (in PyTorch) for some of the major problems in NLP:
   - Text classification, machine translation, question-answering, question generation….

Course Teacher, TAs and Schedule

- Instructor and Coordinator
  - Dr. Tsegaye Misikir Tashu
- TAs:
  - Sophie Sananikone
  - Michal Tešnar
  - Lennart August
- Schedule:
  - Lecture: Tuesday from 17:00-19:00
  - Practical: Thursday from 17:00-19:00

What to Read?

- Speech and Language Processing - Textbook
  - **Dan Jurafsky and James H. Martin (https://web.stanford.edu/~jurafsky/slp3/)**
- Natural Language Processing Conference
  - ACL, NAACL, EACL, EMNLP, CoNLL, Coling, **TACL**
    aclweb.org/anthology
- Machine Learning Conference
  - ICML, NIPS, ECML, AISTATS, ICLR, **JMLR**, **MLJ**
- Artificial Intelligence Conference
  - AAAI, IJCAI, UAI, **JAIR**



Assessment and Grading

- Assignment (3 students per group) (25%)
  - Part 1- Review of Relevant Literature on Empathy Detection and Emotion Classification (12%)
  - Part 2- Implementing an Emotion Detection Classifier (13%)

- Final project (3-4 students per group) (45%)
  - Proposal (5 %)
  - Project report and implementation (30 %)
  - Presentation (10 %)
- Written exam (30%)

- To Pass the course, complete all the assessments and get a 5 in group work (Assignment final project) and the written exam.

Final Project

- Work in groups (3-4 students)
- Project proposal
  - 1-page maximum
- Project report
  - To be submitted before the final presentation
  - The template will be provided
- Project presentation
  - in-class presentation (tentative)
  - Each group will have 10/15 minutes

Upcoming deadline

- Assignment 1 and Assignment 2 are out
- Please enrol into the groups to get access to the assignments
- The submission deadline
  - Task 1 (Assignment 1) will be the **20th of February 2024**
  - Task 2 (Assignment 2) will be the **29th of February 2024**

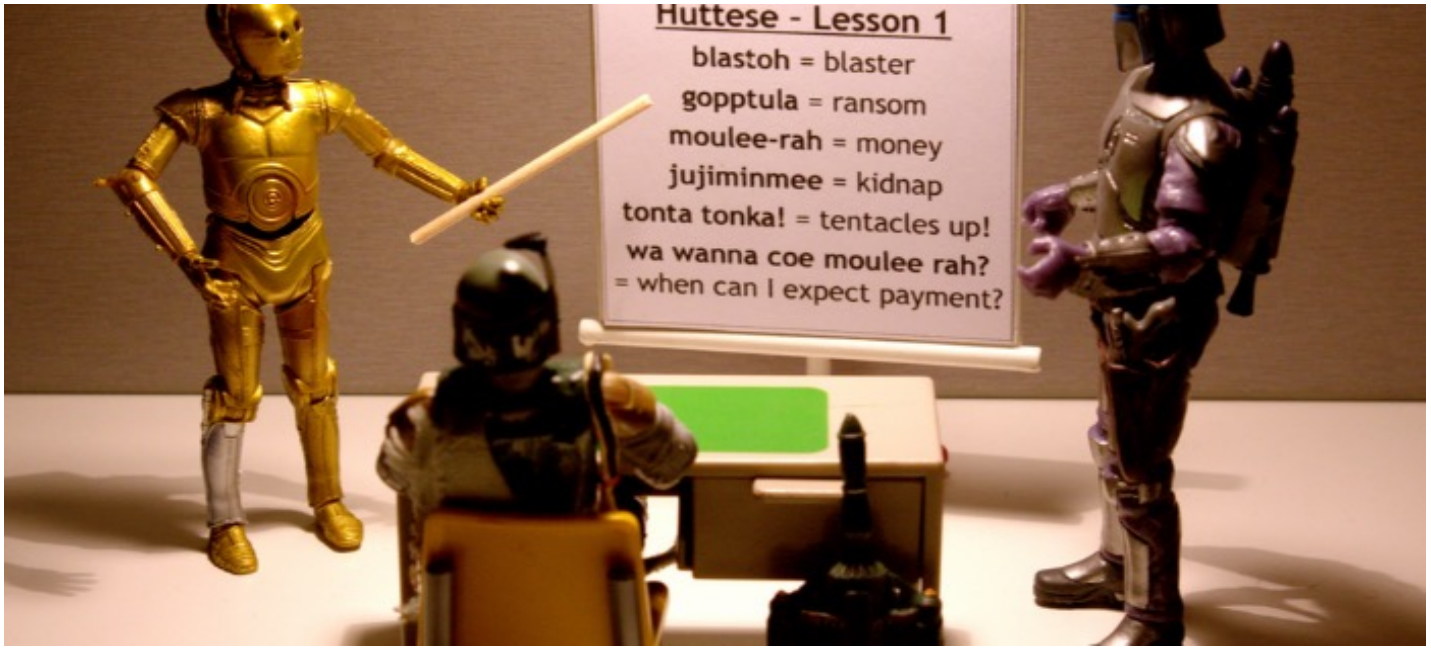
Course Contents

- Introduction to NLP (W1)
- N-gram-based Language Modeling (W1)
- Word Vectors (W2)
- Neural Networks for NLP (W2)
- Recurrent Neural Networks (W4)
- Encoder-Decoder models (W5)
- Attention and Transformers (W6)
- Pretraining and Fine-tuning (W7)

What is NLP?

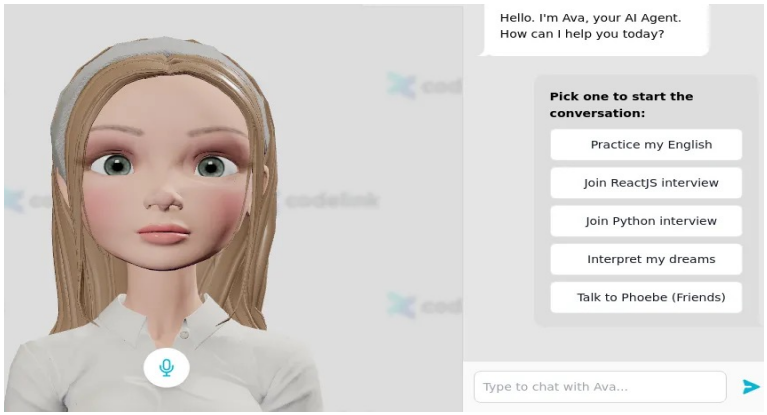
What is NLP

- Focuses on the design and analysis of computational algorithms and representations for processing natural human language [Eisenstien, 2018]



What is NLP?

- Automating the analysis, generation, and acquisition of human (“natural”) language.



Virtual assistant [1]

What NLP topics should I teach BSC students in 2024: Provide me the topics only

ChatGPT

1. Introduction to Natural Language Processing (NLP)
2. Text Preprocessing Techniques
3. Tokenization and Lemmatization
4. Part-of-Speech Tagging
5. Named Entity Recognition
6. Text Classification
7. Sentiment Analysis
8. Topic Modeling
9. Word Embeddings (e.g., Word2Vec, GloVe)
10. Sequence-to-Sequence Models
11. Neural Language Models (e.g., LSTM, Transformer)
12. Attention Mechanisms
13. Transfer Learning in NLP
14. Language Generation (e.g., GPT, BERT)
15. Text Summarization
16. Question Answering Systems

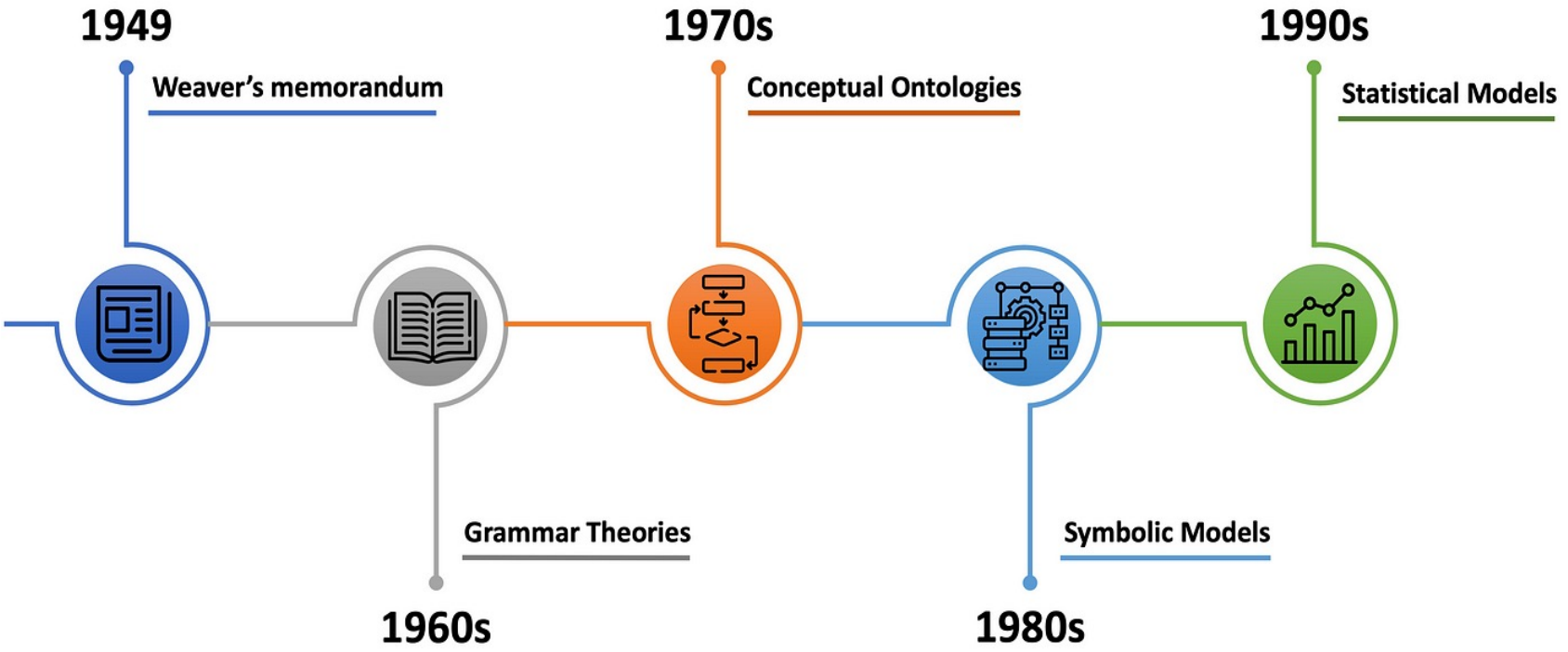
1:Source: https://topai.tools/s/virtual-assistant-avatar-maker

NLP is an interdisciplinary field

- Natural language processing (NLP) combines Computational Linguistics, Machine Learning, and Deep Learning to process human language.

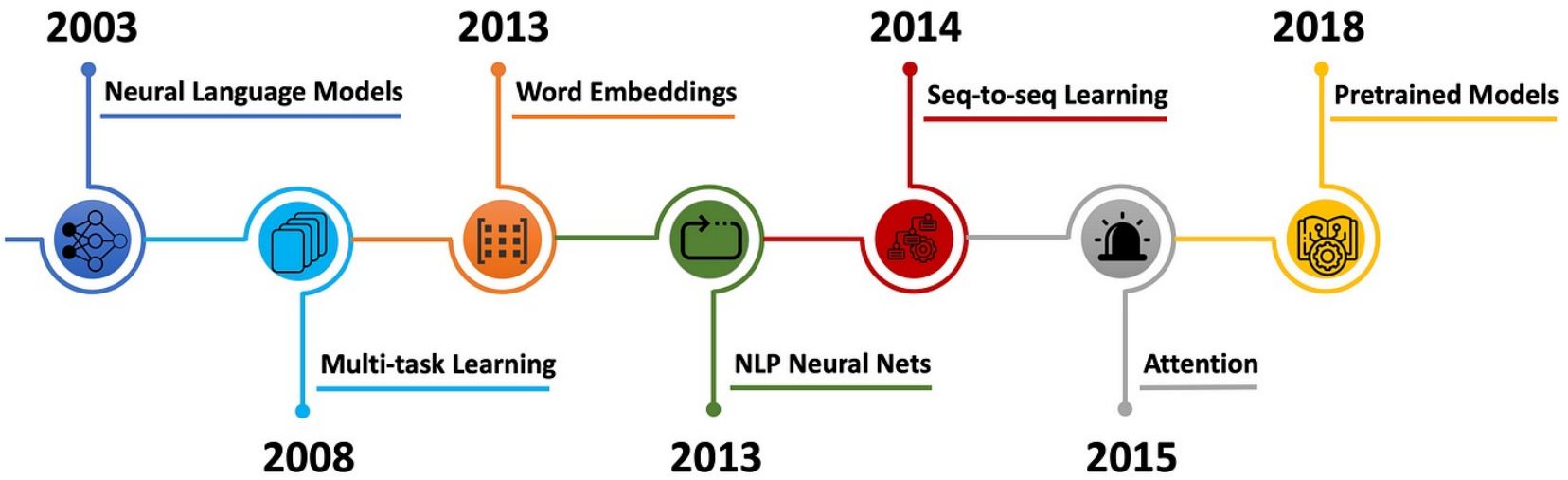
  - Together, these technologies enable computers to process human language in text or voice data and 'understand' its full meaning, complete with the speaker or writer's intent and sentiment.

A Brief History of NLP



Source: https://medium.com/@antoine.louis/a-brief-history-of-natural-language-processing-part-1-ffbcb937ebce

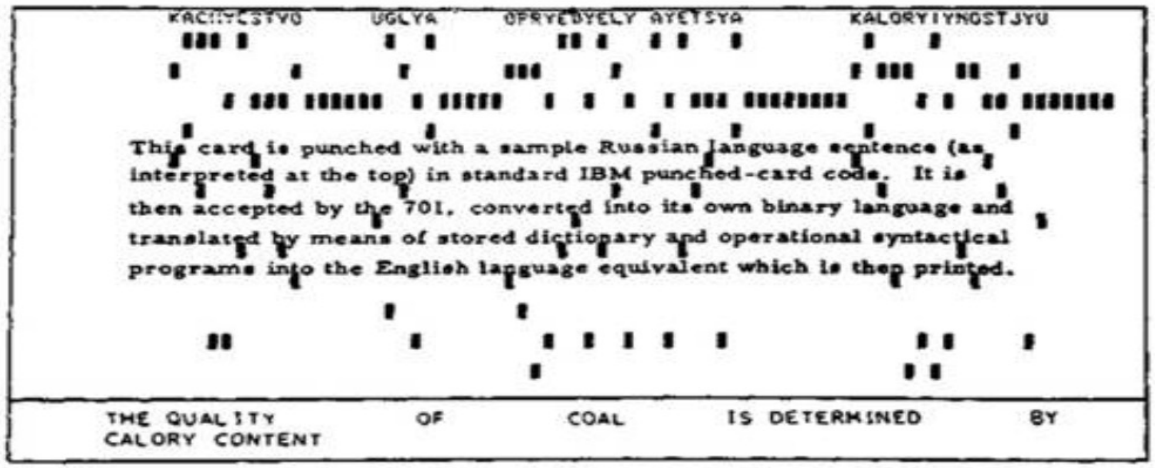
A Brief History of NLP



2018: BERT 2019: T5, RoBERTa 2020: GPT-3

2022: ChatGPT, 2023 GPT-4, Llama, Bard

https://medium.com/@antoine.louis/a-brief-history-of-natural-language-processing-part-2-f5e575e8e37



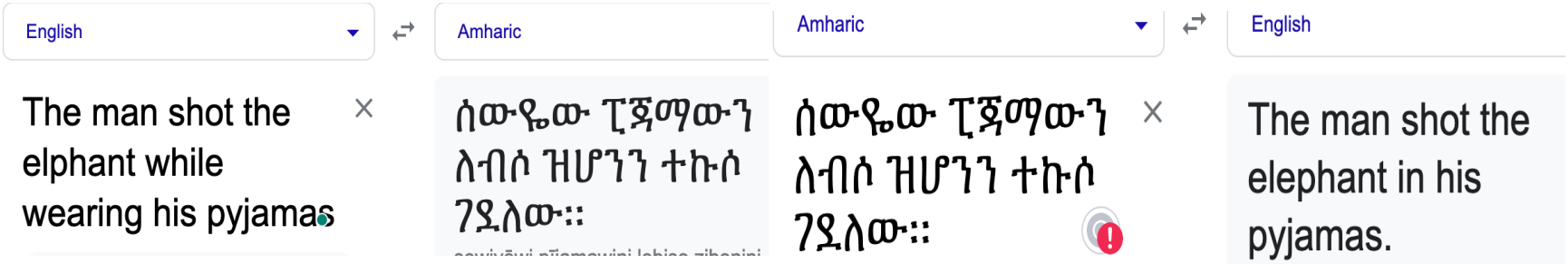
Specimen punched card and below a strip with translation, printed within a few seconds

How it's started

Georgetown experiment 1954.

"Within three or five years, machine translation will be a solved problem"

How it's going?



(Based on Google Translate results in 2024-01-25)

Why is language difficult to understand?

Why It's Hard

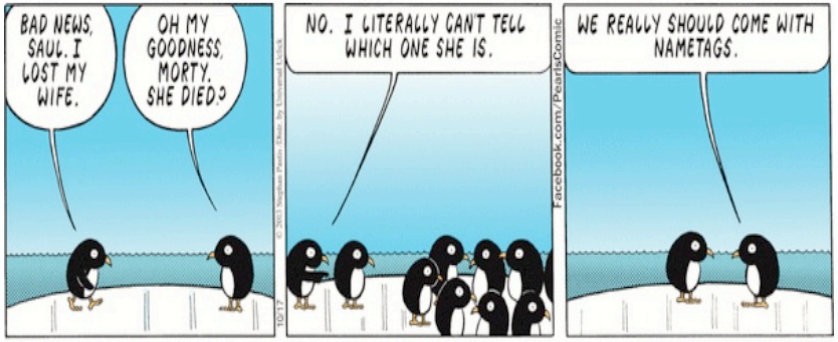
- The mappings between levels of linguistic representation (**Morphology, Syntax, Semantics, Pragmatics**) are extremely complex.
- The appropriateness of a representation depends on the application.
- Ambiguity
- Dialects
- Accents
- Humour, sarcasm, irony
- Context, dependencies

Challenges-Language

- Three challenging properties of language: discreteness, compositionality, and sparseness.
  - **Discreteness**: we cannot infer the relation between two words from the letters they are made of (e.g., hamburger and pizza).
    - For example, the word "bank" can refer to a financial institution or the edge of a river.
  - **Compositionality**: the meaning of a sentence goes beyond the individual meaning of their words.
  - **Sparseness**: The way in which words and discrete symbols can be combined to form meanings is practically infinite.

Challenges – ambiguity

Word sense ambiguity



credit: A. Zwicky

Challenges – ambiguity

Word sense / meaning ambiguity

Credit: http://stuffsirisaid.com

Challenges – language is not static

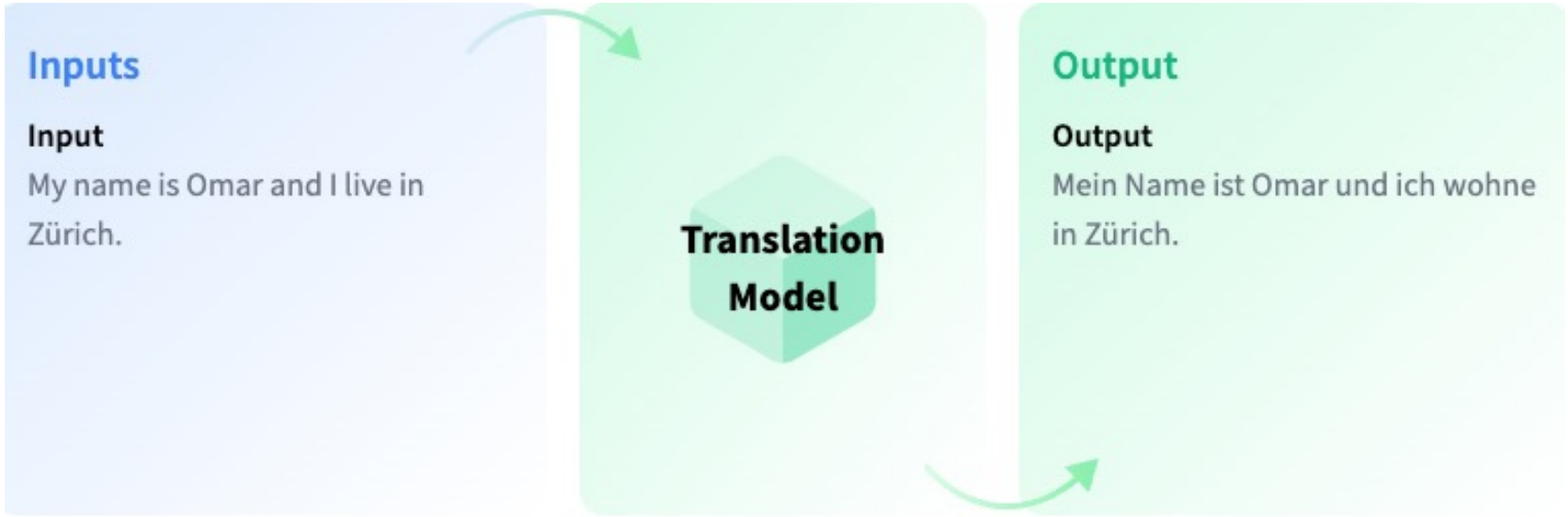
- Language grows and changes
  - e.g., cyber lingo

LOL	
G2G	
BFN	
B4N	
Idk	
BRB	
LYTD	

Key NLP use cases/ Subfields

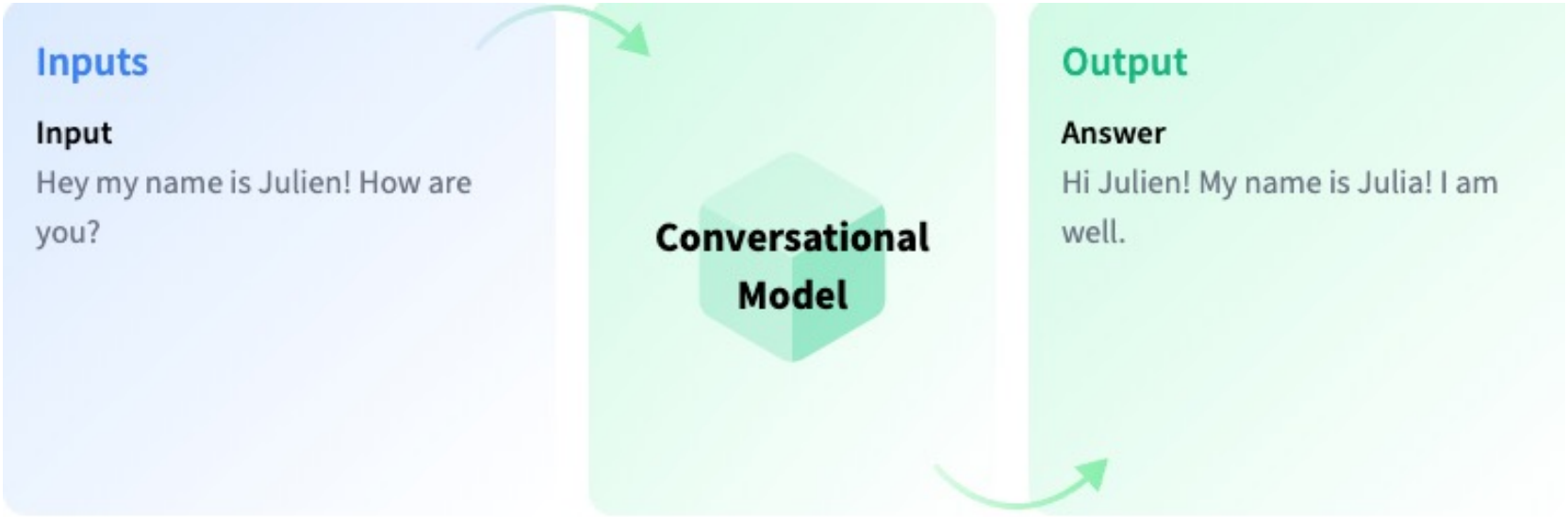
Machine Translation

- Translation converts a sequence of text from one language to another.
- It is one of several tasks you can formulate as a sequence-to-sequence problem



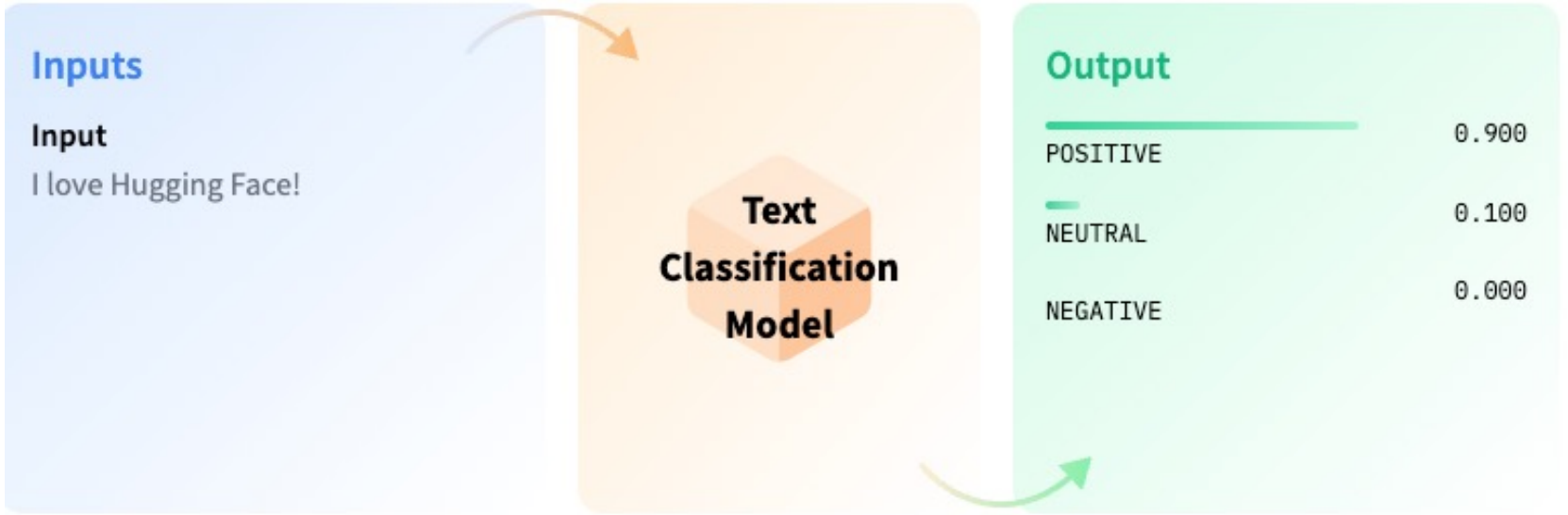
Dialogue/conversational Systems

- The dialogue system is the task of "understanding" natural language inputs - within natural language processing to produce output.
- Designed to simulate human conversation through text, speech, or other modalities.



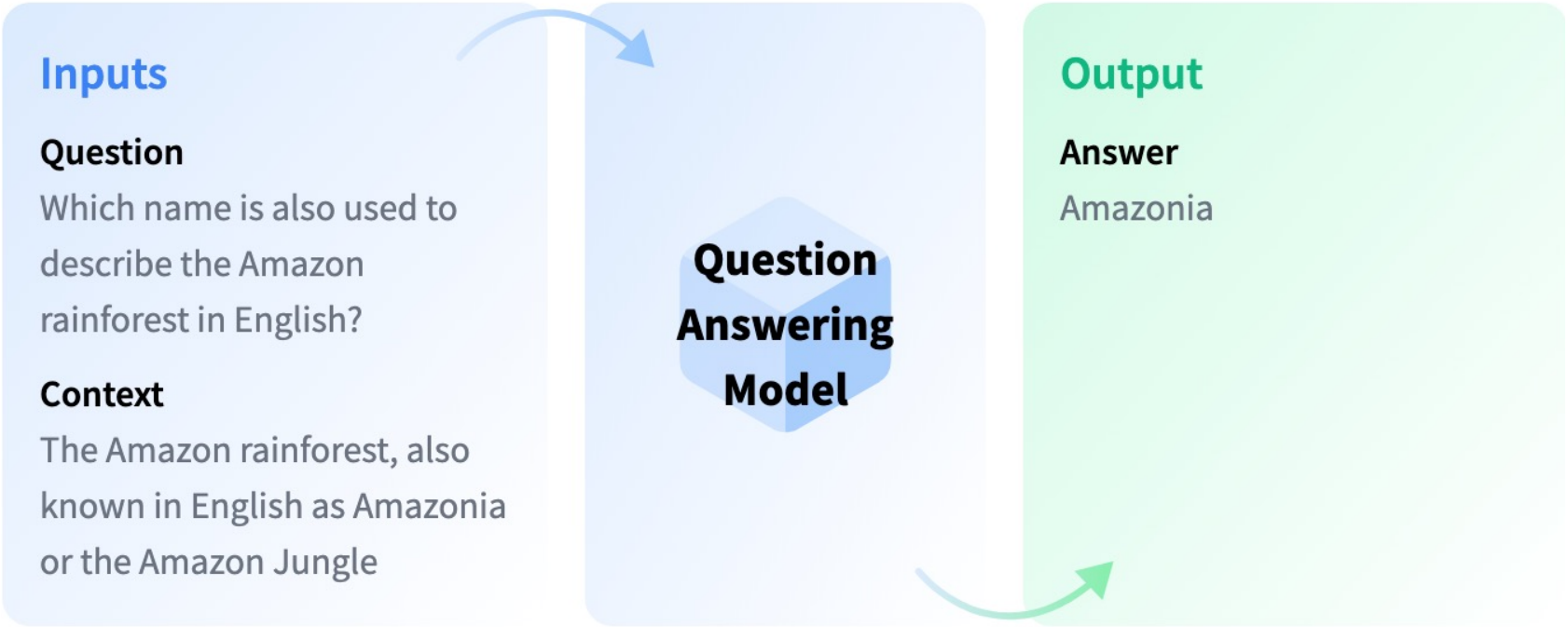
Text Classification

- Text Classification is the task of assigning a label or class to a given text.
- Some use cases are sentiment analysis, natural language inference, and assessing grammatical correctness.



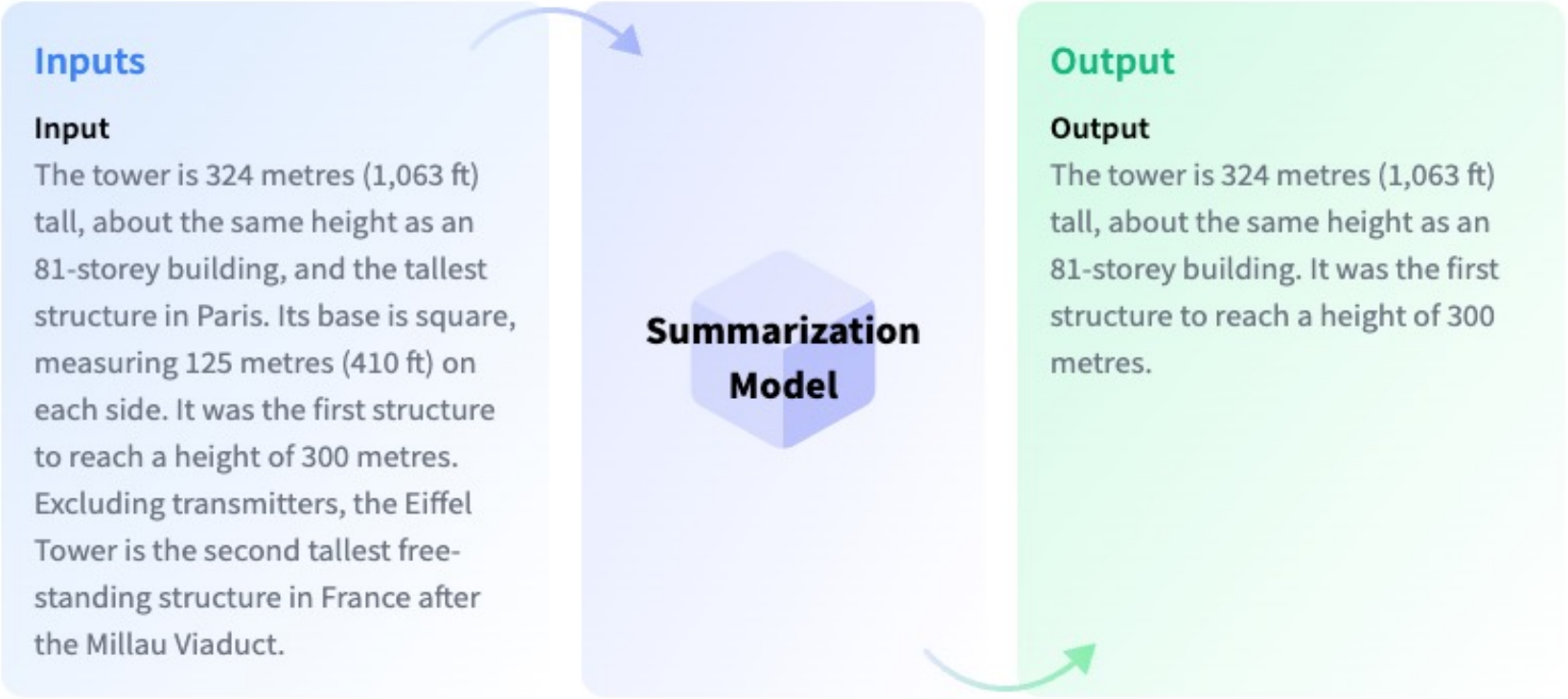
Question Answering (QA)

- QA models can retrieve the answer to a question from a given text, which is useful for searching for an answer in a document.
- Question-answering models can also generate answers without context!



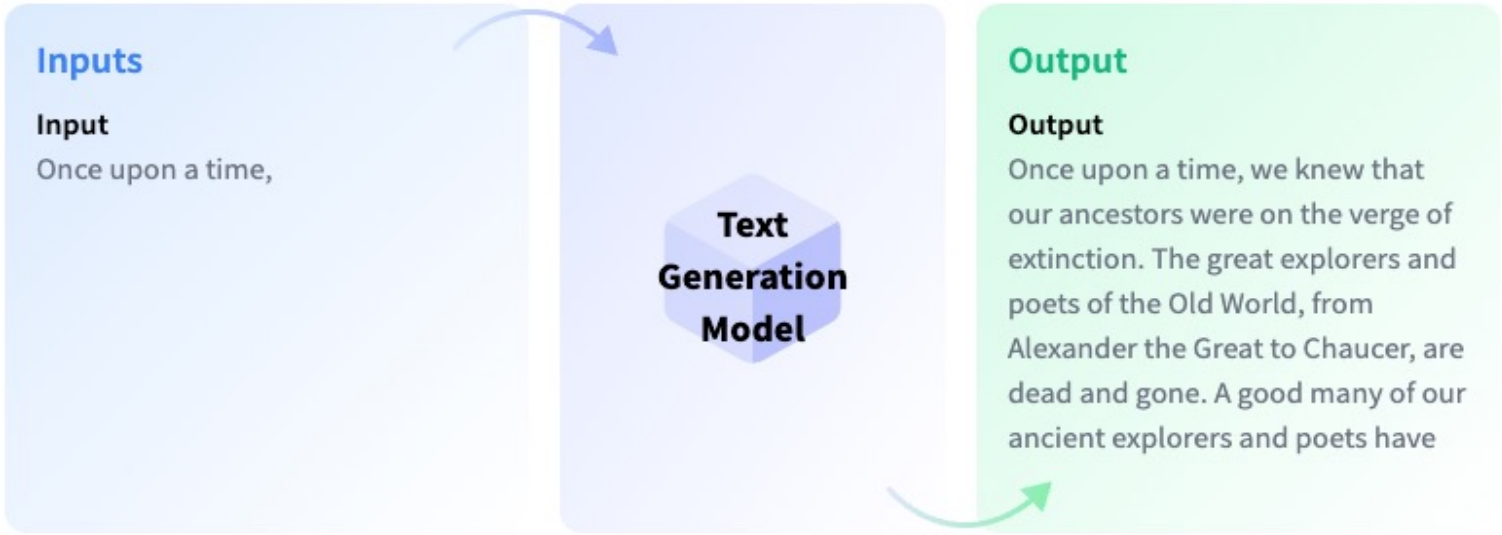
Text Summarization

- Summarization is the task of producing a shorter version of a document while preserving its important information.
- Some models can extract text from the original input, while other models can generate entirely new text.



Text Generation

- Text Generation is the task of producing new text.
- These models can, for example, fill in incomplete text or paraphrase or generate text.



Information Extraction

Is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents.

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.

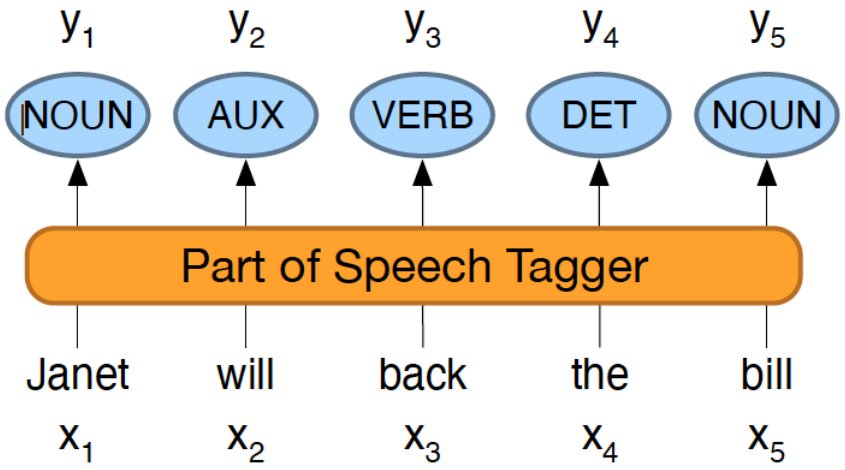
Person	Company	Post
Russell T. Lewis	New York Times newspaper	president and general manager
Russell T. Lewis	New York Times newspaper	executive vice president
Lance R. Primis	New York Times Co.	president and CEO

Key NLP Components

Part of speech tagging (POS)

- POS is the process of assigning part-of-speech tags to each word in a text.
  - The input is a sequence of $x_1,x_2,..,x_n$ of (tokenized) words and a tagset
  - The output is a sequence of $y_1,y_2,\ldots,y_n$ of tags, each output y_i corresponding exactly to one input x_i
- Tagging is a disambiguation task;
  - Words are ambiguous—have more than one possible part of speech—and the goal is to find the correct tag for the situation.
  - For example, book that flight => book can be a verb or
    hand me that book => a noun
- The goal of POS tagging is to resolve these ambiguities, choosing the proper tag for the context.

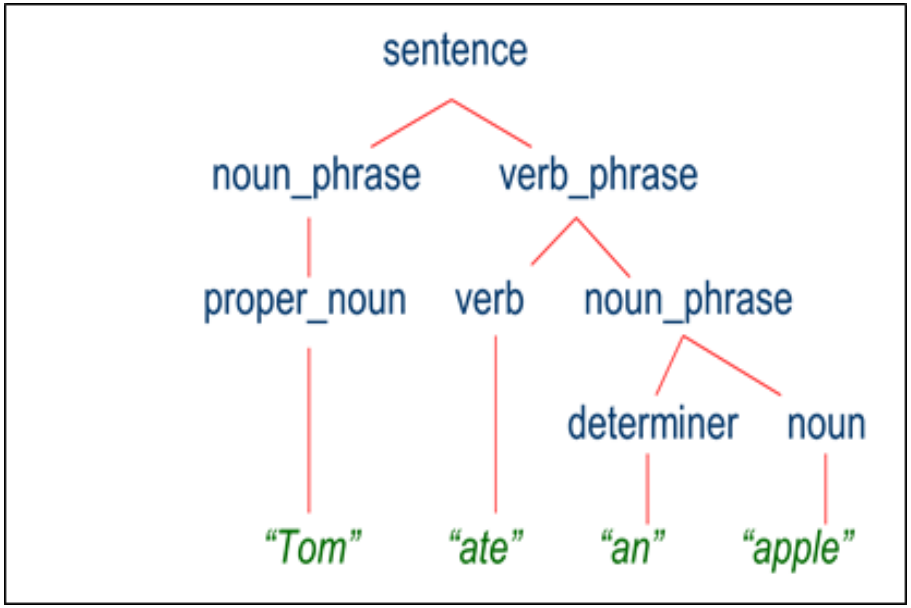
Part-of-speech tags



	Tag	Description	Example
Open Class	**ADJ**	Adjective: noun modifiers describing properties	*red, young, awesome*
	ADV	Adverb: verb modifiers of time, place, manner	*very, slowly, home, yesterday*
	NOUN	words for persons, places, things, etc.	*algorithm, cat, mango, beauty*
	VERB	words for actions and processes	*draw, provide, go*
	PROPN	Proper noun: name of a person, organization, place, etc..	*Regina, IBM, Colorado*
	INTJ	Interjection: exclamation, greeting, yes/no response, etc.	*oh, um, yes, hello*
Closed Class Words	**ADP**	Adposition (Preposition/Postposition): marks a noun's spacial, temporal, or other relation	*in, on, by under*
	AUX	Auxiliary: helping verb marking tense, aspect, mood, etc.,	*can, may, should, are*
	CCONJ	Coordinating Conjunction: joins two phrases/clauses	*and, or, but*
	DET	Determiner: marks noun phrase properties	*a, an, the, this*
	NUM	Numeral	*one, two, first, second*
	PART	Particle: a preposition-like form used together with a verb	*up, down, on, off, in, out, at, by*
	PRON	Pronoun: a shorthand for referring to an entity or event	*she, who, I, others*
	SCONJ	Subordinating Conjunction: joins a main clause with a subordinate clause such as a sentential complement	*that, which*
Other	**PUNCT**	Punctuation	; , ()
	SYM	Symbols like $ or emoji	$, %
	X	Other	asdf, qwfg

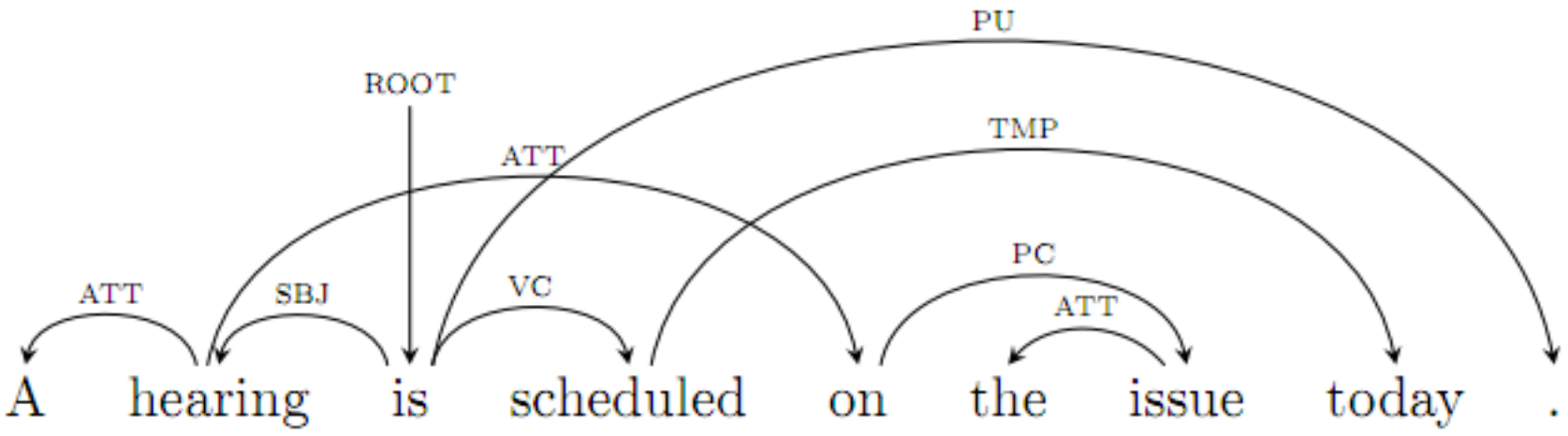
Parsing: Syntactic (Constituency) parsing

- Parsing is the process of determining the syntactic structure of a text by analyzing its constituent words based on an underlying grammar (of the language).
  - Is used to draw exact meaning or dictionary meaning from the text.
  - Comparing the rules of formal grammar, syntax analysis checks the text for meaningfulness.



Dependency Parsing

- In dependency-based approaches to syntax, the structure of a sentence is described in terms of a set of binary relations that hold between the words in a sentence.
- The relations in a dependency structure capture the head-dependent relationship among the words in a sentence.



Co-reference Resolution

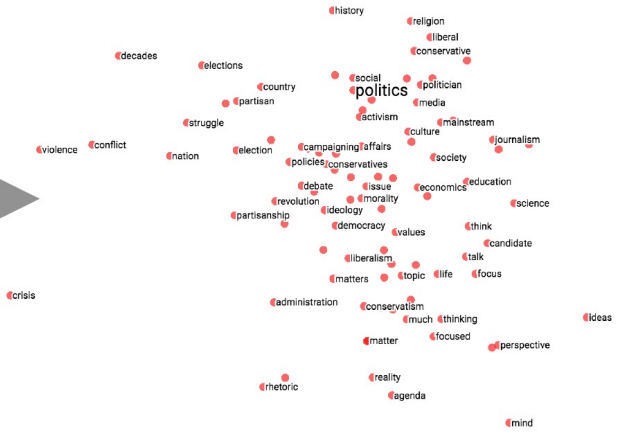
- Coreference resolution is the task of finding all expressions that refer to the same entity or concept in a text.
- Essential for NLP systems to understand the meaning and context of a text. Without it, NLP models would struggle to:
  - Disambiguate pronouns: Which "he" or "she" are you referring to?
  - Track information across sentences
  - Summarize text
  - Answer questions

Christopher Robin is alive and well. **He** is the same person that you read about in the book, **Winnie the Pooh**. As a **boy**, **Chris** lived in a pretty home called **Cotchfield Farm**. When **Chris** was three years old, **his father** wrote a poem about **him**. The poem was printed in a magazine for others to read. **Mr. Robin** then wrote a book.

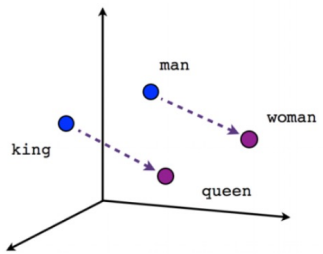
Representation learning-**Distributed representations**

- Learn compact representations of features

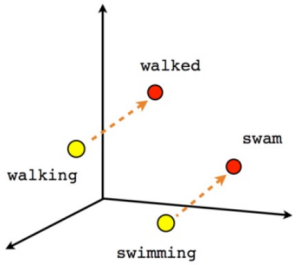
Project words onto a continuous vector space



Similar words closer to each other



Male-Female



Verb tense

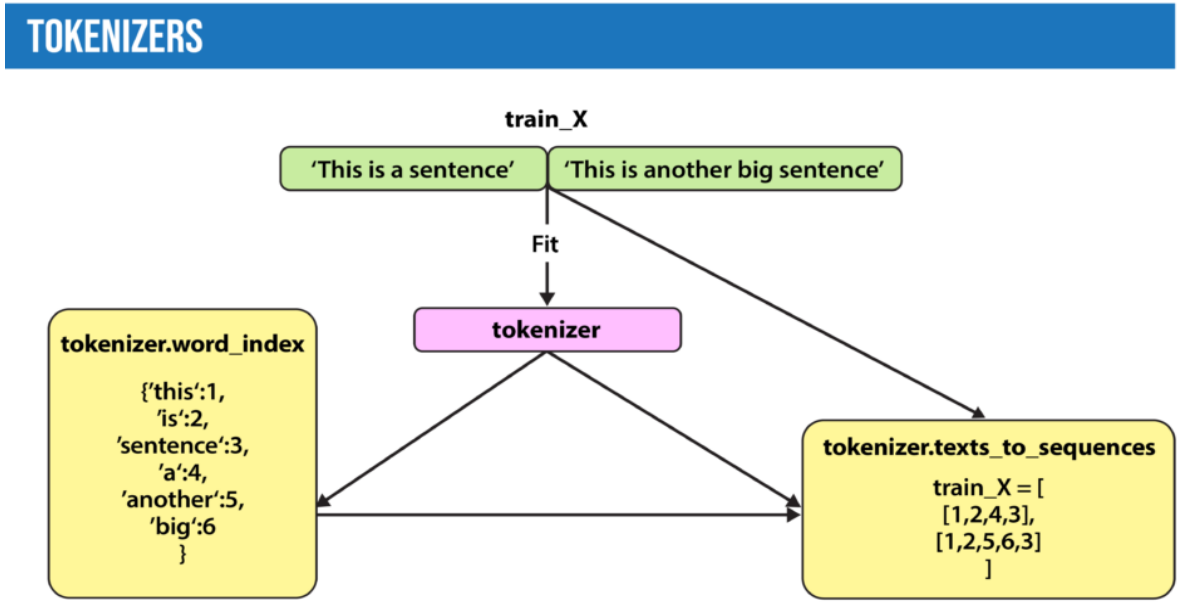
$$v(\text{king}) - v(\text{man}) + v(\text{woman}) = v(\text{queen})$$

How Does Natural Language Processing (NLP) Work?

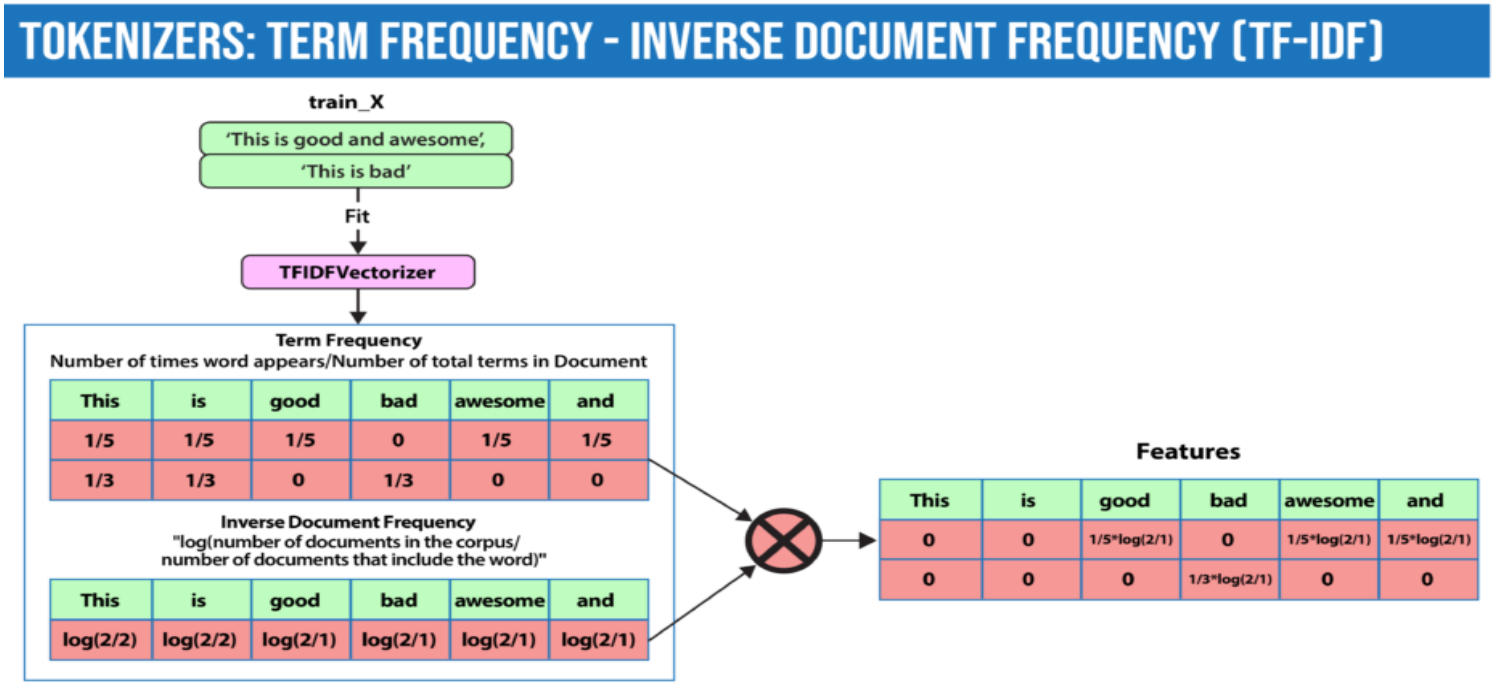
How Does Natural Language Processing (NLP) Work?

- NLP models work by finding relationships/patterns between the constituent parts of the language.
    - for example, the letters, words, and sentences found in a text dataset.
- NLP architectures use various methods for data pre-processing, feature extraction, and modeling.

- **Data pre-processing:** Before a model processes text for a specific task, the text often needs to be pre-processed to improve model performance or to turn words and characters into a format the model can understand/use.
  - Stemming and Lemmatization; Sentence Segmentation
  - Stop word removal; Tokenization

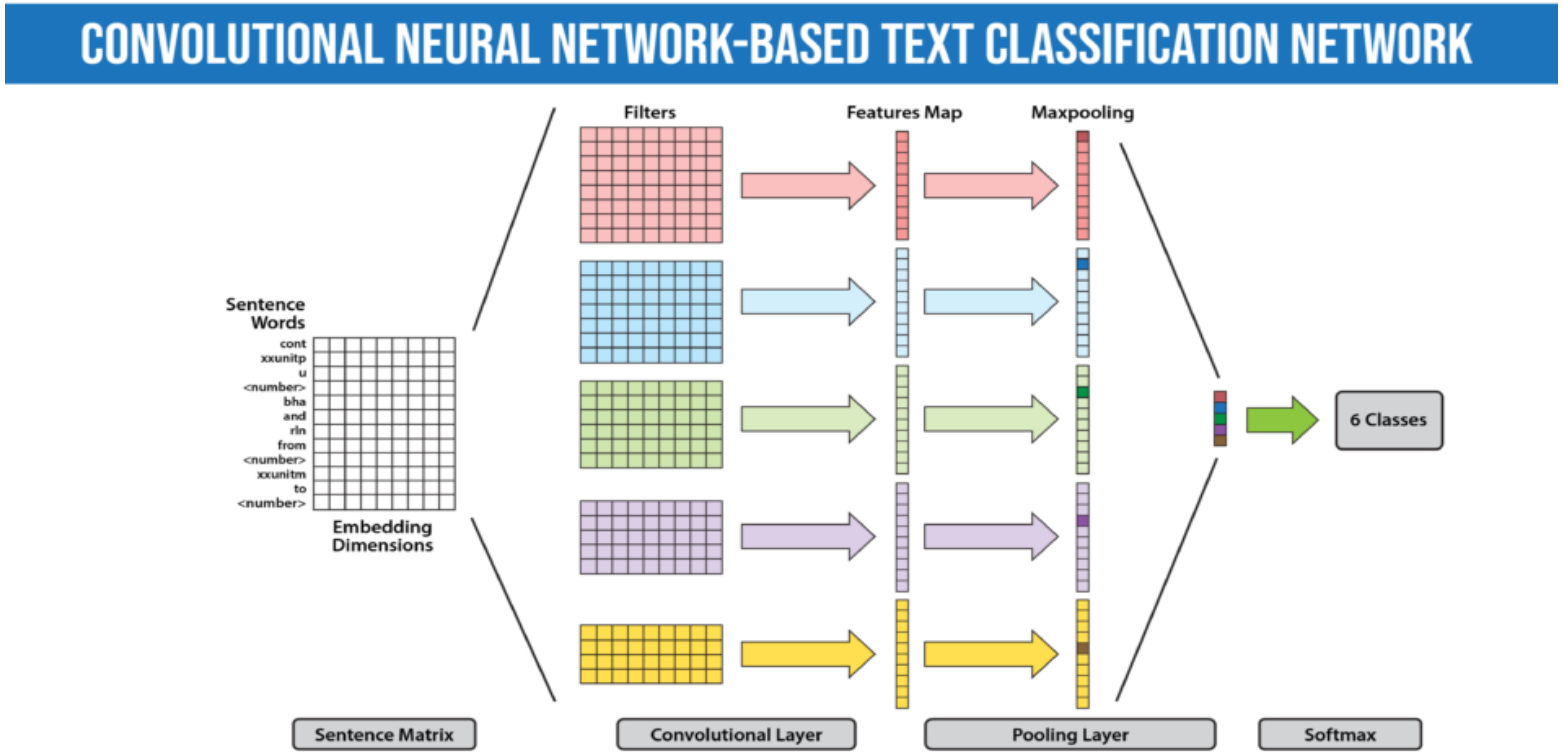


- **Feature Extraction:** Most conventional machine-learning techniques work on the features – generally numbers that describe a document in relation to the corpus that contains it.
  - More popular techniques include Bag-of-Words, TF-IDF Word2Vec, and GLoVe.



TF-IDF creates features for each document based on how often each word shows up in a document versus the entire corpus.

- **Modeling:** After data is pre-processed and feature extracted, it is fed into an NLP architecture that models the data to accomplish a variety of tasks.



Given a sentence, a convolutional neural network uses convolutional layers to refine representations of input words, before combining them to render a classification.

Programming Languages, Libraries, And Frameworks For Natural Language Processing (NLP)

- Natural Language Toolkit (NLTK) is one of the first NLP libraries written in Python.
  - It provides easy-to-use interfaces to corpora and lexical resources such as WordNet.

- SpaCy provides pre-trained word vectors and implements many popular models like BERT.
  - Can be used for building production-ready systems for named entity recognition, part-of-speech tagging, dependency parsing, sentence segmentation, text classification

- Deep Learning libraries: Popular deep learning libraries include TensorFlow and PyTorch
- Hugging Face offers open-source implementations and weights of over 135 state-of-the-art models.

Next lecture: n-gram language models