COE548: LARGE LANGUAGE MODELS

Topic: Natural Language Processing & Human Language

Outline

Recap: LLM Definition

What is NLP?

Human Language

- Lexical and Language Ambiguity
- Language Imprecision and Vagueness

Large Language Models (LLMs)

- **Definition**: An LLM is a computational model capable of language generation or other natural language processing tasks. As language models, LLMs acquire these abilities by learning statistical relationships from vast amounts of text during a self-supervised and semi-supervised training process. (Wikipedia definition)

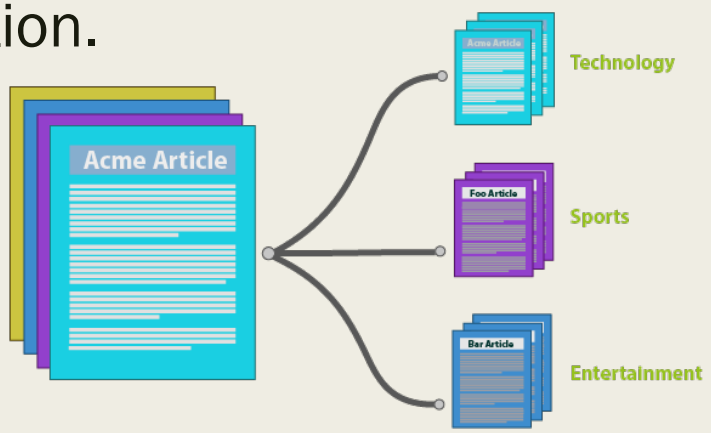
- **Large**: Enormous size of the training data and number of parameters.

- **General Purpose**: The model is sufficient enough to solve commonality of human languages.

Natural Language Processing (NLP)

- Natural language processing (NLP) is an interdisciplinary subfield of computer science - specifically Artificial Intelligence - and linguistics.

- Major tasks in Natural Language Processing are speech recognition, text classification, natural-language understanding, and natural-language generation.

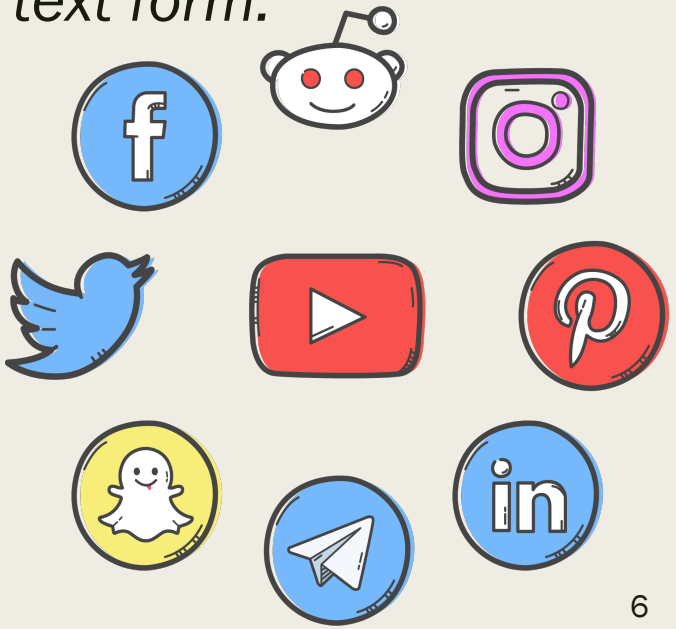


Natural Language Processing (NLP)

- NLP is primarily concerned with providing computers the ability to process data encoded in natural language, typically collected in text corpora, using either rule-based, statistical or neural-based approaches of machine learning and deep learning.

Natural Language Processing (NLP)

- So why do we care about NLP?
- Text is the largest repository of human knowledge.
  - *Wikipedia articles, new articles, scientific articles, patents, social media feeds and posts, etc. all available in text form.*



Human Language

- Human language is a system that was constructed by human beings and somewhat recently (in terms of the age of species on Earth).
  - *It is estimated that human language evolved somewhere in the range of 100,000 – 1,000,000 years ago.*

  - *The power of writing came about some 5,000 years ago – the ability to preserve and share knowledge.*

Human Language

- Human language is a structured (evolving) system which is not so easily modeled. It is rather a complex, rule-governed system with regularities and irregularities.

- Our job is to get better at building computational systems that try to get better at guessing how to convey meaning through words and sentences.

Lexical Ambiguity

- Will Will will Will's will?

- Will (modal verb) Will (name) will (verb) Will's (name) will (noun)?

- Rose rose to put rose roes on her rows of roses.
- Rose (name) rose (verb) to put rose (adj) roes (seafood) on her rows of roses (flowers).

Language Ambiguity (Structural)

- The man saw the boy with the binoculars.
  - *Does "with the binoculars" link to the man or the boy? i.e., Did the man use the binoculars to see the boy, or did the man see a boy using binoculars?*

- Flying planes can be dangerous.
  - *Are the flying planes dangerous or is the action of flying the planes that is dangerous?*

Language Imprecision and Vagueness

- It is very warm here.
  - *Can't tell exact temperature, might mean something different for people from different parts of the world.*

- Q: Did you mother call your aunt last night. A: I'm sure she must have.
  - *"Sure" and "must have" mean probably don't know even though "sure" is usually used for certainty.*

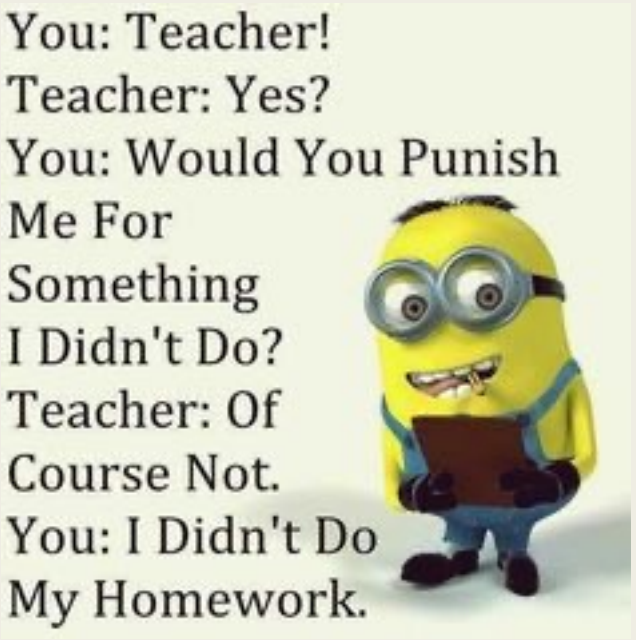
News Headlines

- Hospitals are sued by 7 foot doctors
  - *7 feet tall doctors, or seven doctors that are specialized in feet?*
- Stolen painting found by Tree
  - *Tree might be an organization, but the word "tree" itself would turn this sentence into a very confusing one.*
- Teacher strikes idle kids
  - *Is strikes like hit or grab attention of?*

Jokes

- Why is the teacher wearing sunglasss? Because the class is so bright.

- Why did the bicycle fall over? Because it was two tired.



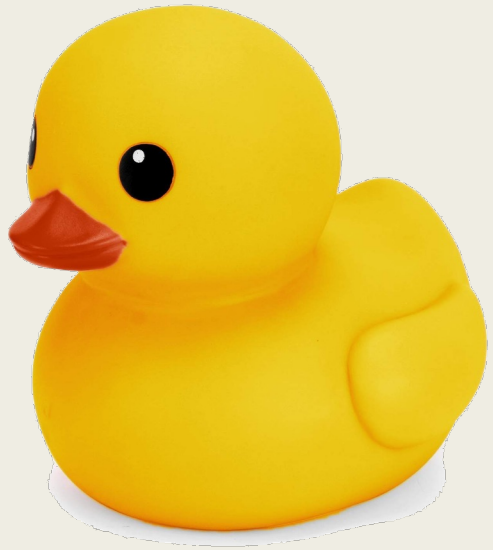
Quick Exercise

- Find 5 meanings for “I made her duck”.

- I cooked a duck for her.
- I cooked a duck belonging to her.
- I made the artificial duck she owns.
- I made her lower her head.
- I waved my magical wand that turned her into a duck?

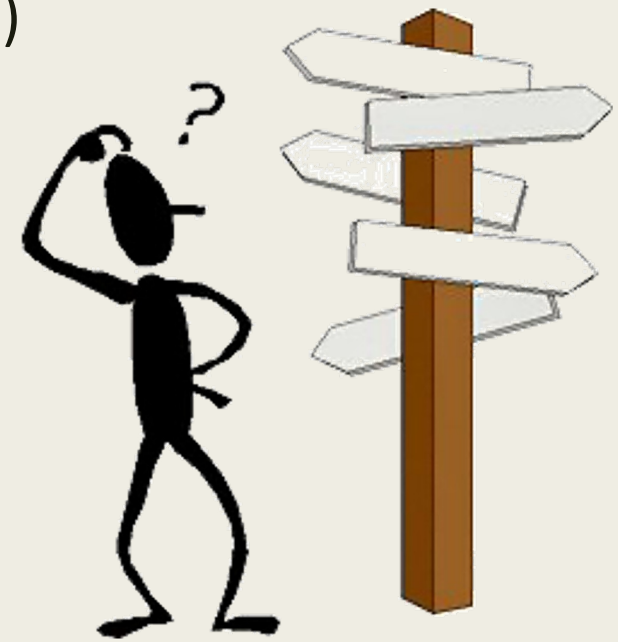
“I made her duck”

- Syntactic category: “duck” can be a noun or verb; “her” can be possessive (“of her”) or dative (“for her”) pronoun.
- Word meaning (“made” can be “cooked” or “created”)
- Grammar (“made” can be transitive, ditransitive, action-transitive
- Phonetics (I’m eight or duck; I’m aid her duck)



More Examples of Ambiguity in Language

- Non-standard English (see you later; c u l8r)
- Segmentation issues (the New York-New Haven Railroad)
- Idioms (dark horse; ball in your court; burn the midnight oil)
- Neologisms (unfriend, retweet, google/skype/photoshop)
- New senses of words (that's *sick* dude)
- Tricky entity names (where is *A bug's Life* playing)



Can we use our advances in AI to process the complex system that language is?

- The question for AI and computer–human interaction is how can we get computers to understand the information conveyed in human languages.
- Fortunately, AI systems may benefit from a virtuous cycle. We need knowledge to understand language and people which is likely contained in the text corpora throughout the world.
- We are now in the era of neural methods and we are now in a space where machine translation works moderately well.

Course Outline

- Begin with revision of linear algebra, calculus, probability, and statistics.
- Rule-based and statistical approaches to NLP, which will cover:
  - *Tokenization*
  - *NLP tasks for understanding language structure*
  - *Word embeddings*
  - *Language models*
- Machine learning/deep learning approaches to NLP → LLMs
  - *Overview of machine learning and deep learning*
  - *Attention mechanisms and transformer models*
  - *Phases of training LLMs*
  - *Etc. (Look at syllabus for full scope of the course)*