Natural Language Processing

Course Overview & Logistics

Ana Marasović
University of Utah

Welcome!

(Professor) Ana (Marasović) [she/her]

- Joined University of Utah in 2022
- *My research:* **NLP/LLMs**, with a focus on **interpretability** and **AI-assisted** *decision making*, *communication*, and *creativity*

Postdoc in **Computer Science** (Seattle, USA)

UNIVERSITY of WASHINGTON

AI2 Allen Institute for AI

PhD in **Comp. Linguistics** (Heidelberg, Germany)

MSc/BSc in **Math** (Zagreb, Croatia)

Grew up here (Omiš, Croatia)

TAs

Fateme Hashemi Chaleshtori [she/her]

3rd year PhD student

She has researched:

- Measuring the usefulness of LLM explanations for improving AI-assisted decision making
- Connecting explanations generated in plain language with internal computations
- Legal NLP

Purbid Bambroo [he/him]

2nd year MS student

Served as a TA for NLP in Spring 2024

He has worked on:

- Supporting research on finding influential train examples for LLM predictions
- Is starting to research RLHF

Overview of Today's Lecture

1. Introduction of Course Staff
2. **What is NLP, Computational Linguistics, AI?**
3. Lecture Organization
    a. Warm-up (5 lectures):

       Practical Intro to Machine/Deep Learning w/ an NLP application
    b. Modern LMs (14 lectures):

       Background, Key ingredients, & In practice
    c. Linguistic structure prediction (6 lectures)
4. Course Logistics

Natural Language Processing (NLP)
...or Computational Linguistics?
......or Machine Learning?
.........or Deep Learning?
............or Generative AI?
...............or AI?

The fundamental question asked in NLP

In what ways can computers understand and use (written forms of) natural language?

- A “natural language” is a language that developed *naturally* in a community
  - Programming languages are artifacts designed deliberately with a goal in mind
  - Ambiguity; No complete solution for formalize a natural language with logic
- Build computer programs that show deep language-understanding & language-use behavior
  - An engineering pursuit that depends heavily on advances in hardware
  - This line of work is usually what people refer to as NLP
- The approach to this today is based end-to-end **deep learning**:
  - *Deep learning:* A sub-field of **machine learning** based on neural networks
  - *Machine learning*: An algorithmic field that designs algorithms that learn to make predictions from data
  - *End-to-end*: No human intervention beyond labeling data & making modeling decisions
  - Specifically, based on **large language models (LLMs)** [general-purpose large-scale neural networks of a particular kind trained on lots of Internet data] which are an example of “**generative AI**”

    ....Is NLP then AI? Is AI then NLP? We’ll come back to this

[Serrano et al., 2023]

The second fundamental question asked in NLP (CL?)

To what extent can the properties of natural languages be simulated computationally?

- NLP x {**linguistics**, cognitive science, psychology}
- How language is structured is an unsolved scientific mystery
- *Scientific lens*:
  - Language is the object of study
  - Experimentally advance the construction of theories about natural language
  - Computational methods play only a supporting role
- *Mathematical lens*:
  - Seeking formal proofs [Ryan Cotterell's lecture; slides 27–48]
- Fundamental linguistic insights gained through this may be crucial for the 1st question
  - This may seem less so for well-served languages, such as English
- This line of work is usually what people refer to as Computational Linguistics today
  - The flagship NLP conference/organization is called Association for Computational Linguistics
  - Some deem NLP and CL to be synonyms; the two fundamental questions are deeply connected

[Serrano et al., 2023]

Discussion: Is NLP then AI? Is AI then NLP?

– History

The phrase "AI": Coined in the **late 1950's** to refer to the aspiration of realizing an **entity possessing the "high-level"** or **"cognitive" capability** of humans to **"reason" & "think"** [Michael I. Jordan]

- **Language processing is** one of the central features of human intelligence, therefore **a prerequisite for "AI"**
- **Reasoning is essential** for basic tasks of **language processing**

Last several decades: **AI ≒ Machine Learning (ML)**

ML experts 🤝 database & distributed-systems experts ⇒ **Data Science**

- The focus shifted from producing entity possessing human-level intelligence to helping make decisions

LLMs like ChatGPT have reignited discussions on human-level intelligence, with significant public attention on a particular notion of "AI" often promoted by "AI doomers" and billionaires, aligned with "TESCREAL ideologies"

This **confluence of ideas and technology trends** has been **rebranded as "AI"** over the past few years

Discussion: Is NLP then AI? Is AI then NLP?

– Answer

It's hard to think about this question due to the messines of the definitions (or lack thereof)

Let's consider take the original focus of AI (an entity that "thinks" and "reasons")

⇒ Language processing is a prerequisite for "AI"
⇒ Reasoning is essential for basic tasks of language processing
⇒ NLP aims to answer the questions beyond the goal of creating such an entity [Emily Bender]
- How are languages similar/different?
- How can we build technology that assists with: transcription, translation, summarization, information access … in different languages?
- How can we evaluate such technology?
- …

⇒ **"NLP intersects with AI, but is neither subsumed by nor subsumes AI"** [Yoav Goldberg]

Recap: What is NLP?

Natural Language Processing (NLP)

A set of methods and algorithms for making natural languages accessible to computers with the goal to build computer programs that show deep language understanding & language-use behavior

Computational Linguistics (CL)

Using computational methods to learn more about how language works

Computational methods based on **machine learning**, **deep learning,** & **LLMs** are ubiquitous in developing such software today, & can be used to assist answering questions about language

This is a common way NLP/CL terms are used, but some view it as synonyms

NLP & AI

"NLP intersects with AI, but is neither subsumed by nor subsumes AI" [Yoav Goldberg]

Definitions [Optiz, Wein, Schneider, 2024]

Linguistics

The study of systematicity and variation in communication between humans, as transmitted via speech, sign, and writing.

Natural Language Processing (NLP)

The field concerned with *developing technology* for sophisticated computational processing of text, and especially, meaning-focused processing of individual sentences, documents, or conversations (as opposed to drawing inferences about entire collections).

Computational Linguistics, narrowly (CL)

The field concerned with computational formalization and processing for the end goal of studying how language works.

Notes on definitions [Optiz, Wein, Schneider, 2024]

Contemporary NLP research is organized around **a paradigm of empirical evaluation of systems on tasks,** whether connected with *general user applications* (like QA, translation, summarization) or more granular and focused on aspects of the language system (like parsing and coreference resolution).

Methods and sub-questions often overlap between CL and NLP, and consequently both can be found at top NLP conferences.

Computational methods based on **machine learning**, **deep learning,** & **LLMs** are ubiquitous in developing such software today, & can be used to assist answering questions about language

But NLP is not just applied ML: Text data is fundamentally discrete; language is compositional, which allows natural language speakers to make "infinite use of finite means"; language is ambiguous, etc

NLP & AI: "NLP intersects with AI, but is neither subsumed by nor subsumes AI" [Yoav Goldberg]

NLP vs. Other applications of ML

- Text data is fundamentally discrete
  - If text is also output, it is not possible to gradually approach an optimal solution
  - Adding small noise to input might result in huge changes in meaning
- The distribution over words resembles that of a power law:

  There will be a few words that are very frequent, and a long tail of words that are rare
- Language is compositional
- NLP models must be robust to observations that do not occur in the training data
- Language is ambiguous

Overview of Today's Lecture

1. Introduction of Course Staff
2. What is NLP, Computational Linguistics, AI?
3. Lecture Organization
    a. **Warm-up (5 lectures):**
       Practical Intro to Machine/Deep Learning w/ an NLP application
    b. Modern LMs (14 lectures):
       Background, Key ingredients, & In practice
    c. Linguistic structure prediction (6 lectures)
4. Course Logistics

Sentiment Classification

This film is interesting as an experiment but **tells no cogent story**. One might feel virtuous for sitting thru it because it touches on so many IMPORTANT issues but it does so **without any discernable motive**. The viewer comes away with **no new perspectives** (unless one comes up with one while one's mind wanders, as it will invariably do during this **pointless film**). **One might better spend one's time staring out a window at a tree growing**.

How to approach building a sentiment classifier with machine learning?

Input x is a string

🤖⁇

This film is interesting as an experiment but **tells no cogent story**. One might feel virtuous for sitting thru it because it touches on so many IMPORTANT issues but it does so **without any discernable motive**. The viewer comes away with **no new perspectives** (unless one comes up with one while one's mind wanders, as it will invariably do during this **pointless film**). **One might better spend one's time staring out a window at a tree growing**.

$$x \to f(x)$$

Feature vector

🤖🩶💙💚💛💜🤎🧡🩵

0.14
1.53
2.12
0.48
4.24
1.94
2.42
1.35

Step 1: Turning raw text into a list of tokens

Splitting a string into a sequence of *tokens*

Tokens: “basic units which need not be decomposed in a subsequent processing” [Webster and Kit, 1992]

```
====Review Text====
This film is interesting as an experiment but tells no cogent story. One might feel virtuous for sitting thru it because
it touches on so many IMPORTANT issues but it does so without any discernable motive. The viewer comes away with no new
perspectives (unless one comes up with one while one's mind wanders, as it will invariably do during this pointless film).
One might better spend one's time staring out a window at a tree growing.


====Tokenized Review Text====
['this', 'film', 'is', 'interesting', 'as', 'an', 'experiment', 'but', 'tells', 'no', 'co', '##gent', 'story', '.', 'one',
'might', 'feel', 'vi', '##rt', '##uous', 'for', 'sitting', 'thru', 'it', 'because', 'it', 'touches', 'on', 'so', 'many',
'important', 'issues', 'but', 'it', 'does', 'so', 'without', 'any', 'disc', '##ern', '##able', 'motive', '.', 'the',
'viewer', 'comes', 'away', 'with', 'no', 'new', 'perspectives', '(', 'unless', 'one', 'comes', 'up', 'with', 'one',
'while', 'one', "'", 's', 'mind', 'wander', '##s', ',', 'as', 'it', 'will', 'invariably', 'do', 'during', 'this',
'pointless', 'film', ')', '.', 'one', 'might', 'better', 'spend', 'one', "'", 's', 'time', 'staring', 'out', 'a',
'window', 'at', 'a', 'tree', 'growing', '.']


====Vocabulary Indices of Tokens====
[101, 2023, 2143, 2003, 5875, 2004, 2019, 7551, 2021, 4136, 2053, 2522, 11461, 2466, 1012, 2028, 2453, 2514, 6819, 5339,
8918, 2005, 3564, 27046, 2009, 2138, 2009, 12817, 2006, 2061, 2116, 2590, 3314, 2021, 2009, 2515, 2061, 2302, 2151, 5860,
11795, 3085, 15793, 1012, 1996, 13972, 3310, 2185, 2007, 2053, 2047, 15251, 1006, 4983, 2028, 3310, 2039, 2007, 2028,
2096, 2028, 1005, 1055, 2568, 17677, 2015, 1010, 2004, 2009, 2097, 26597, 2079, 2076, 2023, 23100, 2143, 1007, 1012, 2028,
2453, 2488, 5247, 2028, 1005, 1055, 2051, 4582, 2041, 1037, 3332, 2012, 1037, 3392, 3652, 1012, 102]
```

Step 2: Mapping each token to a word representation

For each token in our vocab we have an embedding associated with it

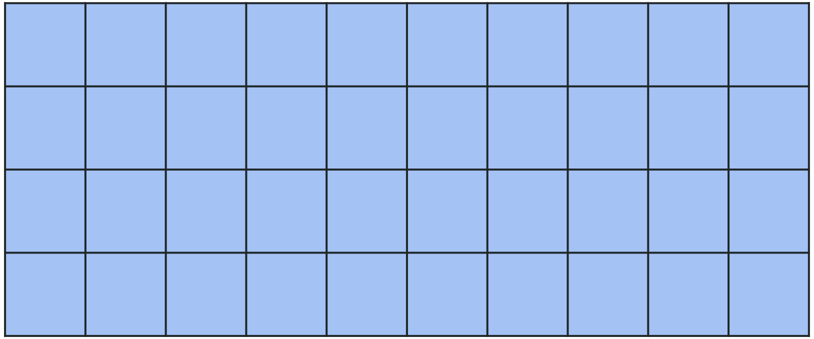
Each row in the “**embedding matrix”** is an embedding/vector of the corresponding token

row 1 = vector of the 1st token in the vocab

row 2 = vector of the 2nd token in the vocab

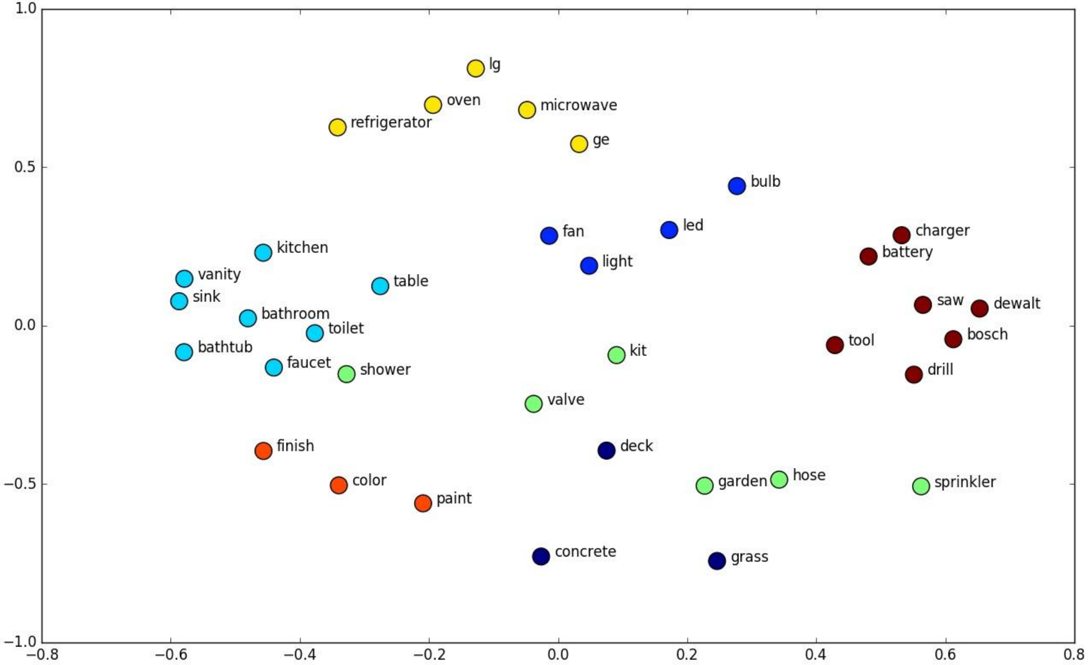
row 3 = vector of the 3rd token in the vocab

row 4 = vector of the 4th token in the vocab



Step 2: Mapping each token to a word representation (cont.)

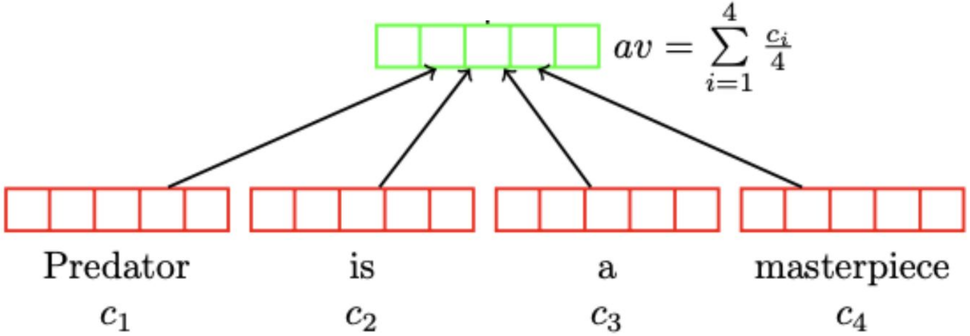
These are not any vectors, but vectors with nice properties such as that words with similar or related means appear close to each other in the vector space



Source: https://neptune.ai/blog/word-embeddings-guide

Step 3: Average word representations

$x \to f(x)$



Input x is a string

🤖⁇

This film is interesting as an experiment but **tells no cogent story**. One might feel virtuous for sitting thru it because it touches on so many IMPORTANT issues but it does so **without any discernable motive**. The viewer comes away with **no new perspectives** (unless one comes up with one while one's mind wanders, as it will invariably do during this **pointless film**). **One might better spend one's time staring out a window at a tree growing**.

$$x \rightarrow f(x)$$

Feature vector

🤖🩶💙💚💛💜🤎🧡🩵

0.14
1.53
2.12
0.48
4.24
1.94
2.42
1.35

Supervised Machine Learning – Training

$$x \rightarrow f(x) \rightarrow m(f(x)) = \hat{y} \in \{+1, -1\}$$

👍 👎

This film is interesting as an experiment but **tells no cogent story**. One might feel virtuous for sitting thru it because it touches on so many IMPORTANT issues but it does so **without any discernable motive**. The viewer comes away with **no new perspectives** (unless one comes up with one while one's mind wanders, as it will invariably do during this **pointless film**). **One might better spend one's time staring out a window at a tree growing**.

0.14
1.53
2.12
0.48
4.24
1.94
2.42
1.35

- We **assume** that the true mapping belongs to some family of functions, e.g., a family of linear functions ($ax + b$)
- **Learn** the coefficients/weights/parameters (a, b) of the function from the **training labeled data**
- **Optimization** problem: Iteratively change the coefficients/weights/parameters until the loss function is minimized
- **Loss function** measures the difference between predicted and actual output values

Supervised Machine Learning – Evaluation

$$x \rightarrow f(x) \rightarrow m(f(x)) = \hat{y} \in \{+1, -1\}$$

This film is interesting as an experiment but **tells no cogent story**. One might feel virtuous for sitting thru it because it touches on so many IMPORTANT issues but it does so **without any discernable motive**. The viewer comes away with **no new perspectives** (unless one comes up with one while one's mind wanders, as it will invariably do during this **pointless film**). **One might better spend one's time staring out a window at a tree growing**.

0.14
1.53
2.12
0.48
4.24
1.94
2.42
1.35

- On the **held-out <u>test</u> instances** we calculate whether the predicted and actual labels are the same (**accuracy**, F1, etc.)
- **Generalization**: Does the mode adapt properly to new, previously unseen data, drawn from the same distribution as the one used to train the model?

Linear classification

$$f(x) \in \mathbb{R}^d \ldots \text{ feature vector}$$

$$w \in \mathbb{R}^d \ldots \text{ weight vector}$$

$$\hat{y} = \begin{cases} +1, w^T f(x) \geq 0 \\ -1, \text{else} \end{cases}$$

Neural classification

Repeat the operations that we have seen before

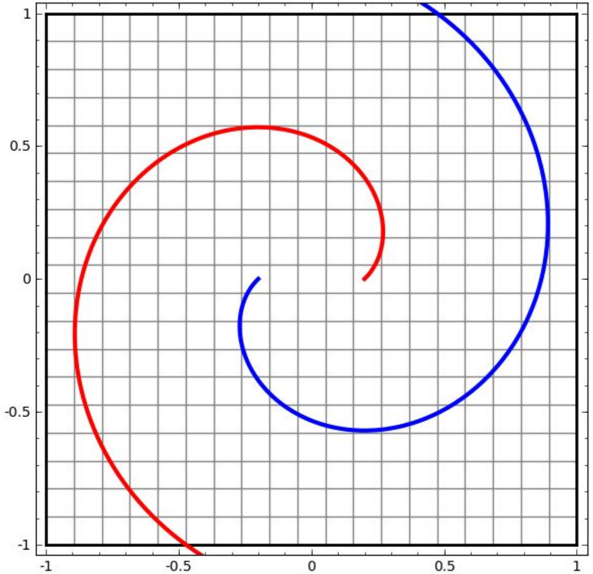
+ apply nonlinear function

$$z_1 = g(W_1 \cdot f(x))$$

$$z_2 = g(W_2 \cdot z_1)$$

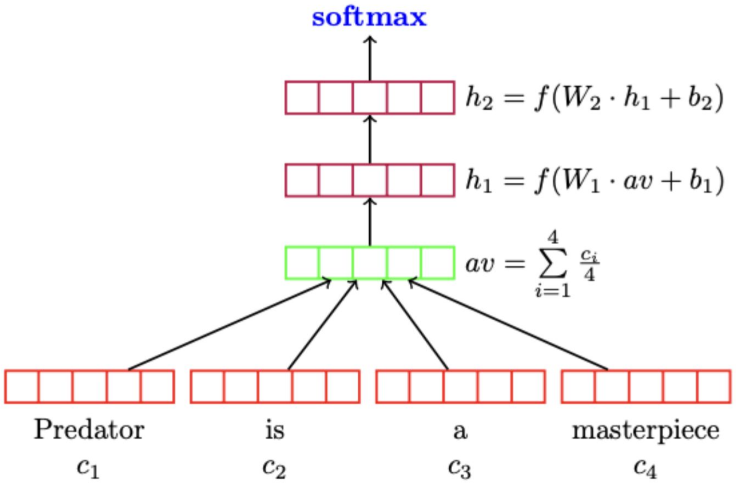
$$\vdots$$

$$z_L = g(W_{L-1} \cdot z_{L-1})$$



Source: https://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

A simple neural network for text classification



Overview of Today's Lecture

1. Introduction of Course Staff
2. What is NLP, Computational Linguistics, AI?
3. Lecture Organization
    a. Warm-up (5 lectures):

       Practical Intro to Machine/Deep Learning w/ an NLP application
    b. **Modern LMs (14 lectures):**

       Background, Key ingredients, & In practice
    c. Linguistic structure prediction (6 lectures)
4. Course Logistics

Modern LMs: **Background** (2 lectures)

Two classic NLP tasks are important for understanding modern LMs:

1. **Language modeling:**

   The task of predicting the next word in a sequence given the sequence of preceding words

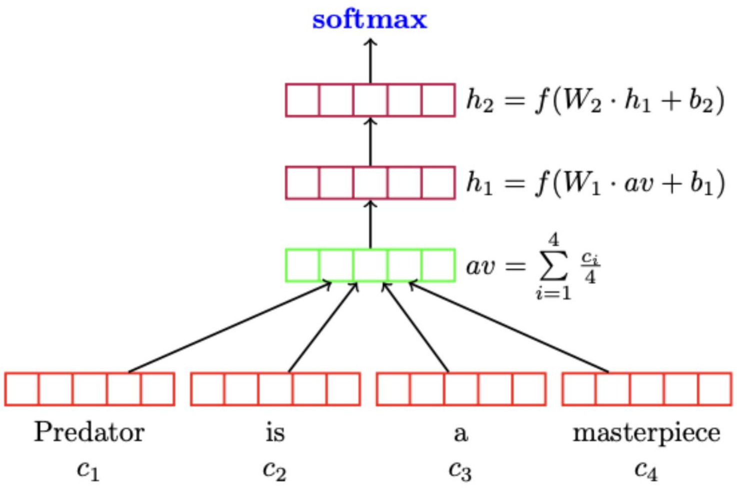
2. **Machine translation:**

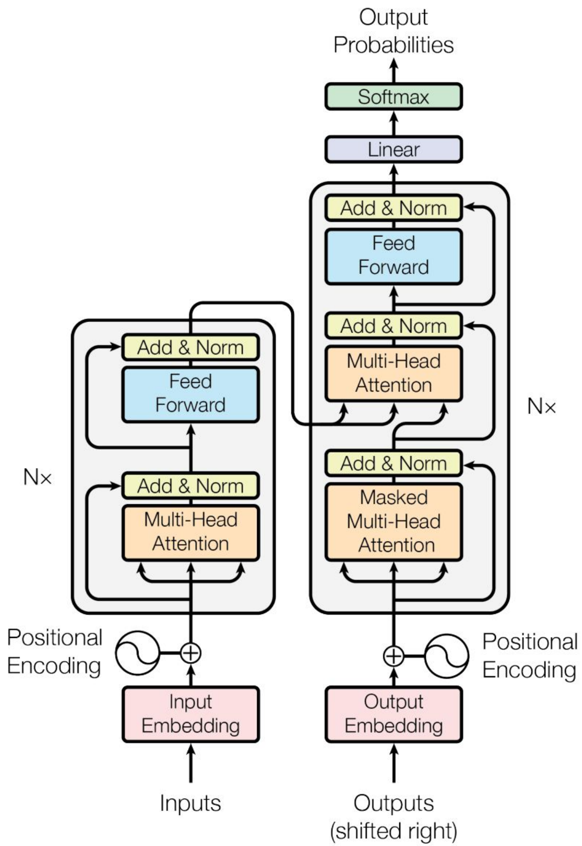
   Translating text from one language to another

Modern LMs: **Key Ingredients** (5 lectures)

1. **Transformer architecture** (2017)
2. **Pretrain-then-finetune paradigm** (2018)
3. **Prompting** (2020)
4. **Learning from human feedback** (2022)

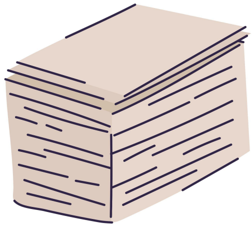
Transformer architecture





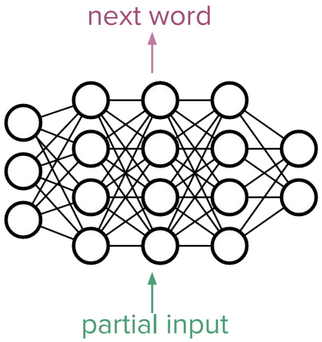
Pretrain-then-finetune

Stage 1:
Pretrain a model



text

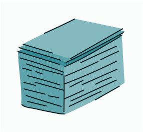
+



neural network
starting from
random weights

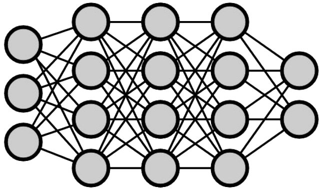
Self-supervised objective: generate next or masked word
(does not require that people label the next word)

Stage 2:
Finetune the model



text + **labels**

+



neural network
starting from
pretrained
weights

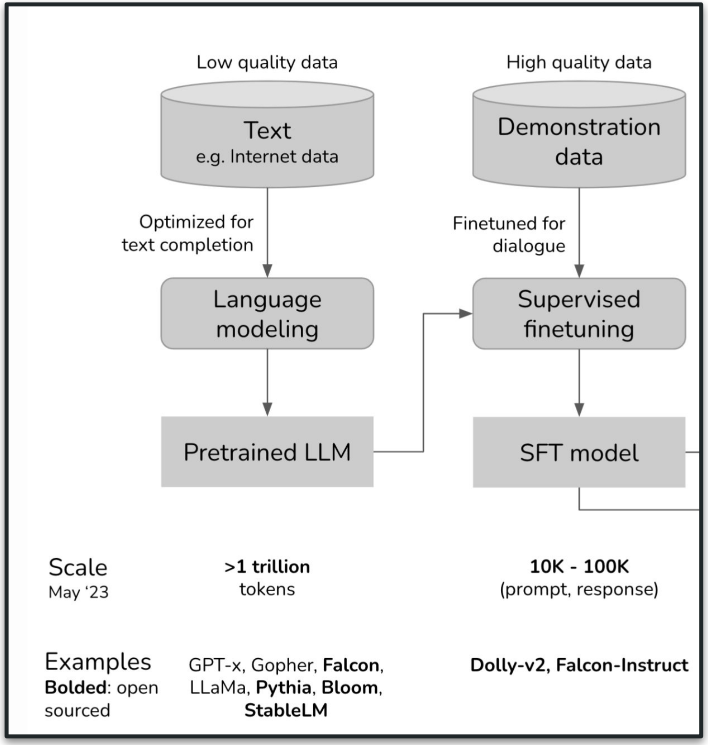
Objective: standard supervised training

Instruction finetuning

A layperson or a domain expert that interacts with an LLM:

- Has no access to model parameters
- Has no knowledge of how to change the model parameters
- But is able to instruct and provide a few examples of their task

We will talk about how to prepare models to successfully **follow any instruction** without require changing model parameters

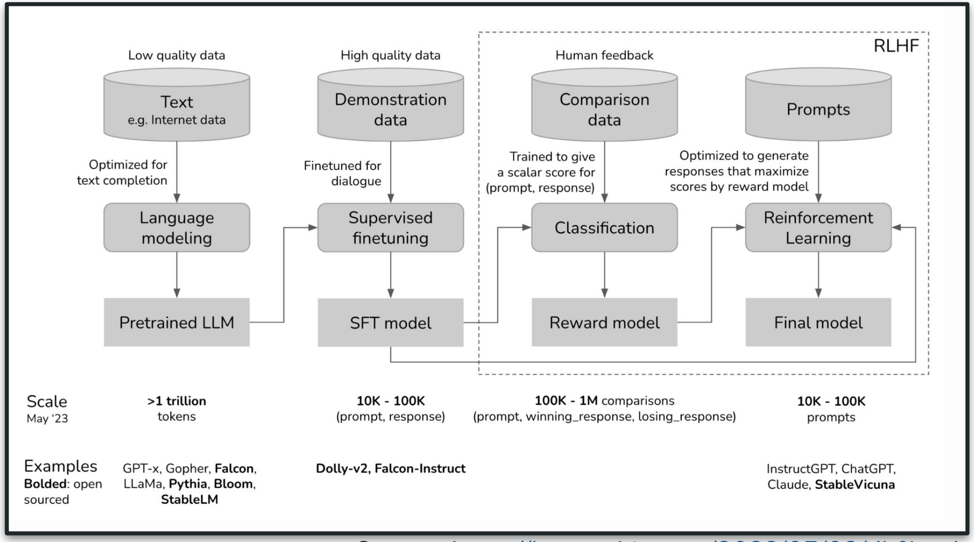


Source: https://huyenchip.com/2023/05/02/rlhf.html

Reinforcement Learning from Human Feedback (RLHF)

General-purpose language technologies like chatbots are **unacceptable** [see Tay fiasco] if it's easy to get **harmful text** from them:

- upsetting user experiences
- enabling harm by aiding violence
- enabling other unlawful activity for a user with malicious intentions



Source: https://huyenchip.com/2023/05/02/rlhf.html

Easier to judge which of 2 texts is safer+useful than to write a safe+useful response for each prompt

Using RLHF, a model learn from these safety/usefulness judgments to produce useful & safe content

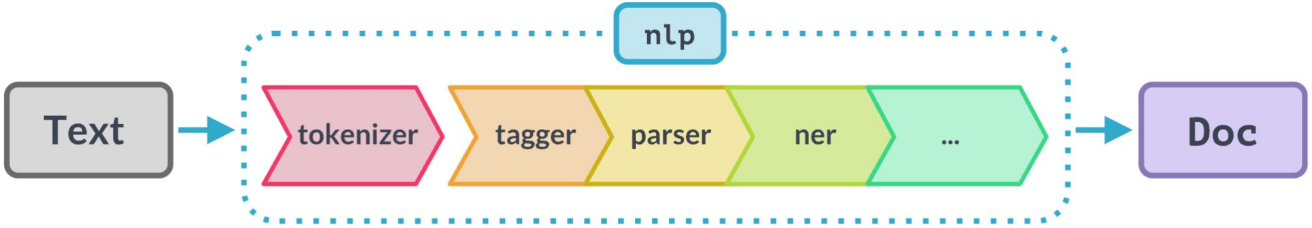
Modern LMs: In Practice (7 lectures)

- **Beyond English text:** Extensions to other languages and vision-and-language problems
- **Efficiency:** Quantization
- **Applications:**
    - *Question Answering* & Retrieval Augmented Generation (RAG)
    - *Summarization* & Text generation evaluation
- **Axes of Modeling & Evaluation Challenges**
- **Responsible deployment of LLMs:** Watermarking and detecting LLM-generated text, supporting privacy, robustness to attacks (or lack thereof), LLM interpretability, preventing amplification of social biases, preventing misuse of LLMs, preventing catastrophic risks, etc.

Overview of Today's Lecture

1. Introduction of Course Staff
2. What is NLP, Computational Linguistics, AI?
3. Lecture Organization
    a. Warm-up (5 lectures):
       Practical Intro to Machine/Deep Learning w/ an NLP application
    b. Modern LMs (14 lectures):
       Background, Key ingredients, & In practice
    c. **Linguistic structure prediction (6 lectures)**
4. Course Logistics

"Standard" NLP Pipeline



NAME	COMPONENT	CREATES	DESCRIPTION
tokenizer	`Tokenizer`	`Doc`	Segment text into tokens.
processing pipeline			
tagger	`Tagger`	`Token.tag`	Assign part-of-speech tags.
parser	`DependencyParser`	`Token.head`, `Token.dep`, `Doc.sents`, `Doc.noun_chunks`	Assign dependency labels.
ner	`EntityRecognizer`	`Doc.ents`, `Token.ent_iob`, `Token.ent_type`	Detect and label named entities.
lemmatizer	`Lemmatizer`	`Token.lemma`	Assign base forms.
textcat	`TextCategorizer`	`Doc.cats`	Assign document labels.
custom	[custom components]	`Doc._.xxx`, `Token._.xxx`, `Span._.xxx`	Assign custom attributes, methods or properties.

Source: https://spacy.io/usage/processing-pipelines

Why should you know classic NLP tasks focused on aspects of the language system? [[Optiz, Wein, Schneider, 2024](#)]

Resources:
Creating lexicons and corpora with sensitivity to language, dialect, genre, & style variations

Evaluation:
Designing human evaluations, interrogating automatic metrics, & analyzing linguistic challenges for systems

Low-resource settings:
Better understand why approaches that work well for English or French might not work well for Swahili or Arapaho

80% languages have no corpus for survived machine learning or LM pretraining

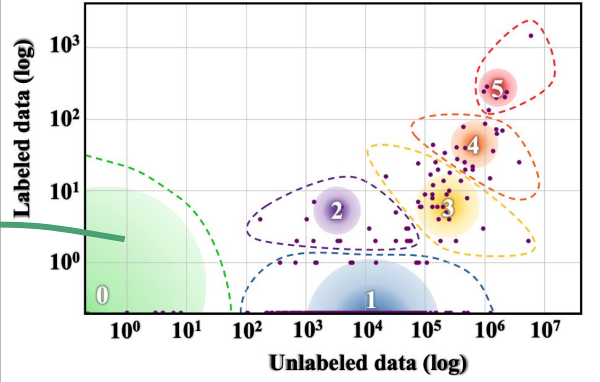
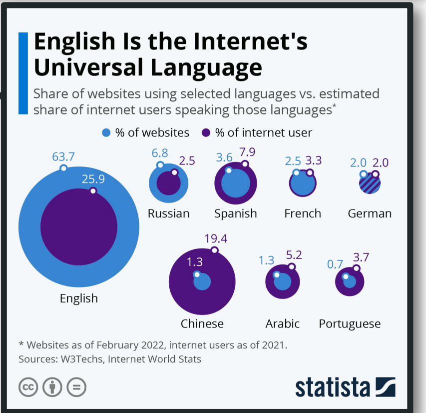


Figure 2: Language Resource Distribution: The size of the gradient circle represents the number of languages in the class. The color spectrum VIBGYOR, represents the total speaker population size from low to high. Bounding curves used to demonstrate covered points by that language class.

"3 in 4 users are unable to understand more than 60% of all websites, at least without a translation tool."



Why should you know classic NLP tasks focused on aspects of the language system? (cont.) [Optiz, Wein, Schneider, 2024]

Interpretability:
We would lack appropriate meta-language for describing many observed patterns

An illustration of this for the question of whether LLMs understand meaning:

- Learning a language: learning how surface forms (language text) connect to underlying structures
- When people and machines do this explicitly ⇒ parsing (part-of-speech tagging, named entity recognition, phrase chunking, coreference resolution)
- How is this core problem addressed in deep learning systems (and in human sentence processing)?

Study of language:
Corpus linguistics, documentary and historical linguistics, …

Lecture schedule is on the class website

https://utah-cs6340-nlp.notion.site

Overview of Today's Lecture

1. Introduction of Course Staff
2. What is NLP, Computational Linguistics, AI?
3. Lecture Organization
    a. Warm-up (5 lectures):

       Practical Intro to Machine/Deep Learning w/ an NLP application
    b. Modern LMs (14 lectures):

       Background, Key ingredients, & In practice
    c. Linguistic structure prediction (6 lectures)
4. **Course Logistics**

CS 5353/6353 (Deep Learning) or **CS 5350/6350** (Machine Learning) are **not formal prerequisites**

But those who completed them will likely more easily understand the class contents and solved the assignments faster

It is possible to take CS 6340/5340 without them, but:

> It would help if you started working on assignments as soon as they are released as you might need more time; check A1 immediately
>
> Proactively seek clarifications during office hours or in Piazza *on time*
>
> Be prepared to read extra materials

Drop deadline is August 30, Friday, not sure about the time

Review this syllabus carefully – students are responsible for understanding everything there!

https://utah.instructure.com/courses/983222

Communication

Join Piazza: https://piazza.com/utah/fall2024/cs5340001fall2024

Issue	Whom to Contact	How
General questions about the content, policies, or assignments	The whole class	Piazza => Public post
Individual questions specific to you incl. requests for accommodations, e.g., individual appointments	Instructor, TAs	Piazza => Private message to the instructor and TAs
Regrade requests	Instructor, TAs	Through Gradescope for those assignments
Urgent accommodations	Instructor	Email ana.marasovic@utah.edu with [CS 6340/5340 F24] at the beginning of the subject line

Communication (cont.)

Join Piazza: https://piazza.com/utah/fall2024/cs5340001fall2024

- We are open to appointments outside of our office hours, but make sure to reach out a few days in advance.
- The instructor and TAs will aim to respond within two business days.
- The instructor will endorse TAs responses if there is nothing else she has to add.
- The instructor and TAs may not read email or check piazza after 5pm.
- The instructor does not read email or check piazza on weekends, or on holidays, and encourages TAs to do the same.
- You can post anonymously to other students on Piazza, but your name will always be visible to the instructor and TAs.
- Feel free to answer questions but do not post potential homework answers.

Office hours

Let’s find out what times in the week might be most useful for the greatest number of students

https://www.when2meet.com/?25946072-QGKr4

Please don’t submit if you doubt you’ll stick with this course

Evaluation

5 Assignments = **70%**

- **A1:** Logistic regression with bag-of-word features 🌶️
- **A2:** Deep averaging network (DAN) with and without pretrained word embeddings 🌶️ 🌶️
- **A3:** Transformer from scratch. Next word prediction 🌶️ 🌶️ 🌶️
- **A4:** Standard finetuning and QLoRA. Decoding strategies. Generation evaluation. 🚧👷🏗️
- **A5**: Various prompting. Model and KV-cache quantization. Jailbreaking. 🚧👷🏗️

Mid-term exam = **15%** (in person; check that you are in SLC on the date)

Final exam = **15%** (in person; check that you are in SLC on the date)

(no projects)

The lowest-scoring assignment for undergraduate students will be dropped

I do not plan to round or curve the grades

No additional chances for earning extra credit

On code review

We expect that you are **experienced with programming in Python**

An important note on code review during office hours:

TAs will follow my guidelines which emphasize not to look at and debug the code directly, but have students go over their approach and lead in the right direction through a conversation. Otherwise, this quickly turns into giving away solutions which is unfair, doesn't promote a deeper understanding, and doesn't enhance problem-solving skills. Machine learning debugging is notoriously challenging. To avoid difficulties, start your assignments as soon as you receive them.

Late submissions

- You are allowed to be late with at most 2 assignments no more than 48H, without penalty. You don't need to ask for permission.
- After 2 assignments have been submitted late, once the deadline is missed, those turned in late will lose 10%. Every next 24 hours until it is turned another 10% is deducted.
- Assignments will not be accepted more than 48 hours late and will be given a 0.
- Assignments will be posted far enough ahead of time that I will not be able to make additional exceptions if a student falls ill. The exception will be prolonged illness accompanied by a doctor's note.

Academic misconduct

Read carefully "Submitting Assignments" in the Syllabus:

https://utah.instructure.com/courses/983222

Don't put yourself in this situation, please

Review this syllabus carefully – students are responsible for understanding everything there!

https://utah.instructure.com/courses/983222