What comes next?

Antoine Bosselut

Announcements: Class

- **Last lecture today!**
- **In-depth Course Evaluation**
  - Please fill it out! Due June 11th!
- **Course Project**
  - Milestone 2 due **Sunday, June 4th!**
  - Final Report due **Sunday, June 18th!**
    - Spec for final report released tomorrow along with guide for getting more Cloud Credits
  - Fill out prior knowledge survey you received by e-mail!

Announcements: External

- **Fall 2023 AE Position in Analysis 1**
  - **Goal**: Develop an automated QA system using LLMs
  - **Contact**: sacha.friedli@epfl.ch
- **EPFL launching MAKE team around large language models and generative AI**
  - Dedicated resources to training large language models at scale (A100 GPUs)
  - Opportunity to complete Masters semester projects and theses on training large models at scale
  - Express interest here **(non-binding)**: https://forms.gle/oRkP2ttT2j82w2VT9

Natural Language Processing

Enabling Human-Machine Collaboration	Accelerating Human-Human Communication	Mining Human Insights
Search Engines	Machine Translation	Sentiment Analysis
Dialogue Agents	Text Summarization	Motivation Analysis
Text Generation	Information Extraction	Emotion Detection

Core Methods

Word Embeddings

- **Words and other tokens become vectors; no longer discrete symbols!**
  - Need to define a vocabulary of words (or token types) *V* that our system can assign to a vector
- Word embeddings can be learned in a self-supervised manner from large quantities of raw text
  - Learning word embeddings from scratch using labeled data for a task is data-inefficient!
- **Three main algorithms:** Continuous Bag of Words (CBOW), Skip-gram, and GloVe

Recurrent Neural Networks

- Early neural LMs (and n-gram models) suffered from **fixed context windows**
- Recurrent neural networks can **theoretically** learn to model an **unbounded context length**
  - no increase in model size because weights are shared across time steps
- Practically, however, **vanishing gradients** stop vanilla RNNs from learning useful **long-range dependencies**
- LSTMs and GRUs are variants of recurrent networks that mitigate the vanishing gradient problem
  - used for for **many sequence-to-sequence tasks**

Transformers

- **Temporal Bottleneck**: **Vanishing gradients** stop many RNN architectures from learning **long-range dependencies**

- **Parallelisation Bottleneck:** RNN states depend on previous time step hidden state, so must be **computed in series**

- **Attention**: Direct connections between output states and inputs (solves temporal bottleneck)
  - **Self-Attention**: Remove recurrence over input, allowing parallel computation for encoding

- **Transformers** use self-attention to encode sequences, but now require position embeddings to capture sequence order

Text Generation

- Text generation is the foundation of many useful NLP applications (e.g., translation, summarisation, dialogue systems)
- **Autoregressive**: models generate one token a time, using the context and previously generated tokens as inputs to generate the next token
- **Teacher forcing** is the premier algorithm for training text generators
- Need better approaches for **automatically evaluating** NLG systems

Modern Methods

Pretraining

- **Contextual representations:** Let us model words and sequences conditioned on the context around them
- **ELMo:** Based on bidirectional LSTMs. **Good for pretrained embeddings.**
- **GPT**: Uses a transformer decoder. **Good for generating text as a language model.**
- **BERT**: Uses a transformer encoder. **Good for classification and sequence labelling.**
- **BART + T5:** Pretraining sequence-to-sequence transformer models. **Extendable to all task types!**

Scale & Compression

- Scale is necessary to achieve many emergent breakthroughs: in-context learning, chain-of-thought, etc.
- Training at scale is very expensive: **months of training —> millions of $$$$**
- Scaling laws estimate optimal model / dataset sizes for compute budget
- Once a large model is trained, compression (size-performance trade-off) can help optimise:
  - Number of parameters (storage space)
  - Inference time
- Different compression techniques
  - Pruning, quantization, factorization, weight sharing, knowledge distillation

Retrieval Augmentation

- Retrieval-Augmented LMs infuse knowledge from external sources into LMs.
  - Suitable for knowledge-intensive tasks where factual accuracy is needed.
- Using external knowledge reduces how much capacity the model uses to memorize information
  - make them smaller in size without compromising performance.
- Recently, models retrieve from tools and APIs to enhance their capabilities

Multimodality

- Language is just “an imperfect, incomplete, and low-bandwidth serialization” of real life
  - Multimodal is the next step!
- Huge amount of multimodal data available, just waiting to be used to train a model.
- The largest LLMs are already LVLMs (GPT-4, PALM-E)
- Many new challenges emerge when dealing with multimodal content!

Challenges

Robustness

- Pretrained models fine-tuned on downstream tasks achieve **incredible performance** on benchmarks designed to measure language understanding
- Benchmarks are made up of datasets that are human-constructed
- Humans make “mistakes” when designing datasets, allowing models to shortcut true understanding of the task in favour of easily learnable heuristics
- Designing challenging evaluations remains a primary goal natural language processing systems

Ethics

- Many ethical considerations go into designing NLP systems
  - **What biases may these systems encode?**
  - **What groups will they exclude?**
  - **Will they produce toxic or misinformed content?**
  - **What private information can they leak about their data subjects?**
  - **Will their interactions with humans open up new avenues for misuse or conflict?**
- Practitioners need to understand these issues to design safer systems for the benefit of all
  - **Sometimes, that can mean not designing a system in the first place!**

Interpretability

- Model interpretability methods allow us to explore *how* our models function when they make predictions
- **Many different approaches**: probing, local explanation, model-based explanation, and many more!
- Interpretability methods can be misleading in what they show — always be aware of the assumptions baked into your interpretation
- **Plausibility** is not the same as **faithfulness** — is this actually how your model is making a prediction? Or do you just want to believe it is?

Opportunities

Education

- Personalized learning for students
  - Adaptive learning platforms that cater to needs and preferences of individual students
  - **Tutoring**: suggest relevant learning resources tailored to students' interests, promoting self-directed learning
  - **Feedback**: personalized feedback helps students identify specific areas where they need to focus and improve.
  - **Engagement**: self-directed learning could enhance student engagement by making learning experience interactive and enjoyable
- Efficient course administration
  - Automate administrative tasks, freeing up valuable time for educators to focus on instruction and student support.
  - **Analytics & Insights**: Analyze vast amounts of educational data to help educators identify areas of improvement, predict student performance, and refine teaching
  - **Grading**: AI algorithms can evaluate and provide feedback on assignments, quizzes, and tests, reducing time educators spend on grading and enabling faster feedback

From **Learning Management Systems** to **Intelligent Tutoring Systems**

Medicine

- **Notable Benchmark**: Med-PaLM 2 model achieved **86.5%** performance on U.S. Medical Licensing Exam–style questions (USMLE)

- On par with "expert" test takers

Example of USMLE-style question

A 32-year-old woman comes to the physician because of fatigue, breast tenderness, increased urinary frequency, and intermittent nausea for 2 weeks. Her last menstrual period was 7 weeks ago. She has a history of a seizure disorder treated with carbamazepine. Physical examination shows no abnormalities. A urine pregnancy test is positive. The child is at greatest risk of developing which of the following complications?

A. Renal dysplasia

B. Meningocele

C. Sensorineural hearing loss

D. Vaginal clear cell carcinoma

Legal

Efficient Legal Administration

- **Legal research and analysis**: search and analyze databases, statutes, regulations, case law, etc.
- **Contract analysis**: review and extract key information from contracts, flag risks, etc.
- **Improved legal outcomes**: assist in preparing cases, identifying precedents, predicting outcomes, etc.

Improved Access to Justice

- Bridge justice gap by providing legal assistance to individuals who may not have access to legal help
- **Virtual legal assistants**: provide basic legal guidance, answer frequently asked questions, assist with legal form completion, etc.

Entertainment

- **Improved content creation and curation**: Enable content creators to generate and curate media more efficiently, enhancing storytelling, visual effects, and audio enhancements.

- **Data-driven decision making**: Analytics tools allow entertainment companies to gather and analyze vast amounts of data, enabling them to make data-driven decisions on content production, distribution, and marketing strategies.

Life comes at you fast!

Legal

- Microsoft CoPilot built on top of OpenAI's Codex, which used code from online repositories as training data

- Microsoft CoPilot shown to reproduce exact code snippets from copyrighted code

MICHAEL A. JACOBS (SBN 111664)
MJacobs@mofo.com
JOSEPH C. GRATZ (SBN 240676)
JGratz@mofo.com
TIFFANY CHEUNG (SBN 211497)
TCheung@mofo.com
MORRISON & FOERSTER LLP
425 Market Street
San Francisco, California 94105-2482
Telephone: (415) 268-7000
Facsimile: (415) 268-7522
[CAPTION PAGE CONTINUED ON NEXT PAGE]

Attorneys for Defendants OPENAI, INC., a Delaware nonprofit corporation, OPENAI, L.P., a Delaware limited partnership, OPENAI GP, L.L.C., a Delaware limited liability company, OPENAI STARTUP FUND GP I, L.L.C., a Delaware limited liability company, OPENAI STARTUP FUND I, L.P., a Delaware limited partnership, OPENAI STARTUP FUND MANAGEMENT, LLC, a Delaware limited liability company

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN FRANCISCO DIVISION

J. DOE 1 and J. DOE 2, individually and on behalf of all others similarly situated, Plaintiffs, v. GITHUB, INC., a Delaware corporation; MICROSOFT CORPORATION, a Washington corporation; OPENAI, INC., a Delaware nonprofit corporation; OPENAI, L.P., a Delaware limited partnership; OPENAI GP, L.L.C., a Delaware limited liability company; OPENAI STARTUP FUND GP I, L.L.C., a Delaware limited liability company; OPENAI STARTUP FUND I, L.P., a Delaware limited partnership; OPENAI STARTUP FUND MANAGEMENT, LLC, a Delaware limited liability company, Defendants.	Case No. 4:22-cv-06823-JST 4:22-cv-07074-JST Hon. Jon S. Tigar **CLASS ACTION** **DEFENDANTS OPENAI, INC., OPENAI, L.P., OPENAI GP, L.L.C., OPENAI STARTUP FUND GP I, L.L.C., OPENAI STARTUP FUND I, L.P. AND OPENAI STARTUP FUND MANAGEMENT, LLC'S NOTICE OF MOTION AND MOTION TO DISMISS COMPLAINT; MEMORANDUM OF POINTS AND AUTHORITIES** Date: May 4, 2023 Time: 2:00 p.m. Courtroom: 6

Public Opinion

State of the field

- Advances in hardware and algorithms allowed NLP to make rapid progress
- Still lots of exciting work to be done in the field!
- **Now that the systems work,** we must be clear on how to safely and reliably deploy systems

Thanks for a great semester!

Guest Lecturers

Gail Weiss

Mohammadreza Banaei

Angelika Romanou

Negar Foroutan

Syrielle Montariol

Thanks for joining us!

Antoine Bosselut
Instructor

Deniz Bayazit
TA

Angelika Romanou
TA

Molly Petersen
TA

Mete Ismayil
AE

Zeming Chen
TA

Silin Gao
TA

Mohammadreza Banaei
TA

Simin (Olivia) Fan
TA

Sepideh Mamooler
AE

NLP @ EPFL

- **Natural Language Processing** Lab
  - Master's Theses, Semester Projects available every term
  - MAKE project: kicking off this Fall!
- Other **NLP** courses
  - **Fall 2023**: Introduction to Natural Language Processing (6 credits)
    - Lectures, Quizzes, Final Exam
  - **Fall 2023**: Topics in Natural Language Processing (2 credits)
    - Paper reading, paper reviewing, discussion