ACL 2023 notes

ACL acceptance rate: 20.73%
Long: 22.13%
Short: ~15%
Demos: ~37%
Submission countries: 1. China 2. USA
Presidential address: ACL 2024 will be ARR only

Best and Outstanding Papers:
https://2023.aclweb.org/program/best_papers/

Prof. Hajič in Outstanding Papers:

What's the Meaning of Superhuman Performance in Today's NLU?

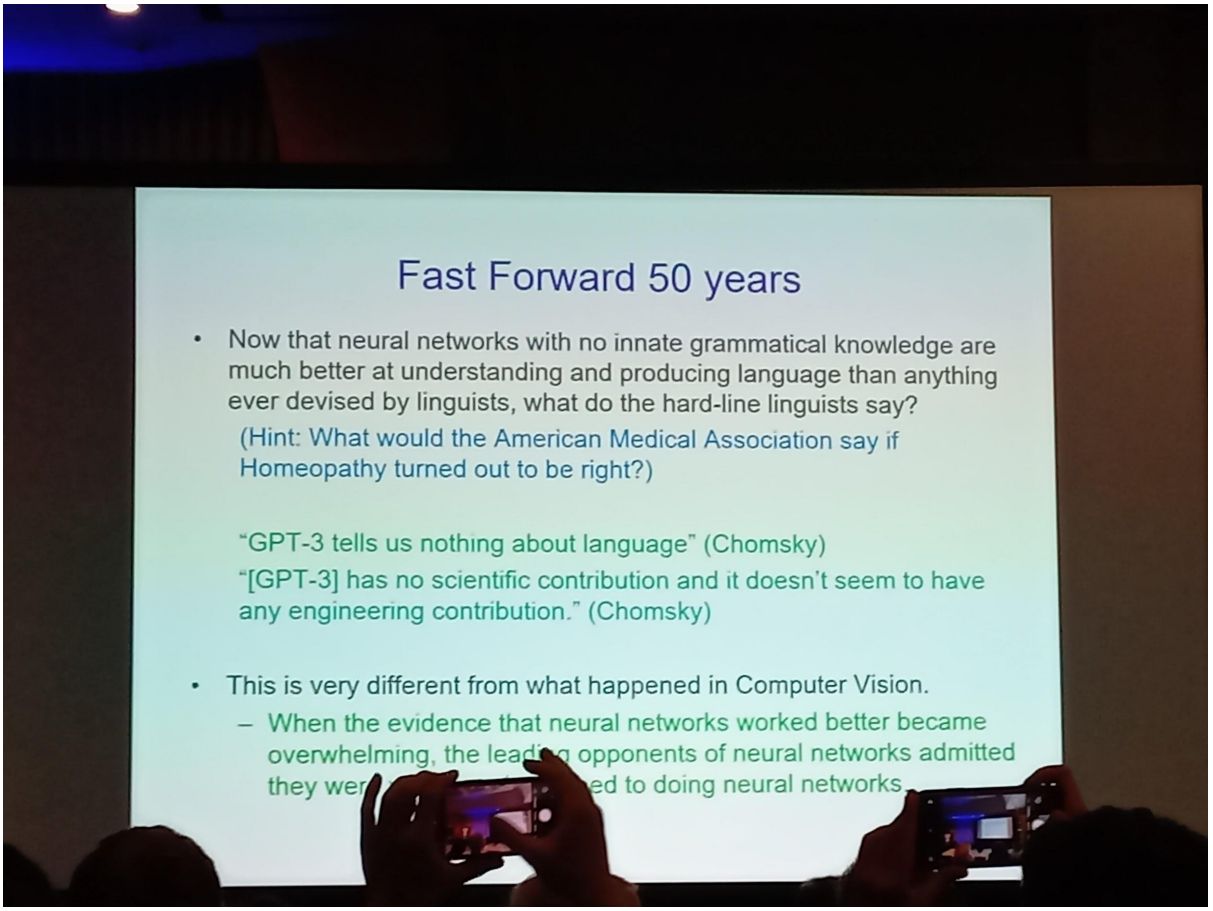
Simone Tedeschi, Johan Bos, Thierry Declerck, Jan Hajič, Daniel Hershcovich, Eduard Hovy, Alexander Koller, Simon Krek, Steven Schockaert, Rico Sennrich, Ekaterina Shutova and Roberto Navigli

https://aclanthology.org/2023.acl-long.697/

Keynote: Hinton

Early NN limitation: could not do recursion
How to do true recursion with NN?

Fast forward 50 years: people say NNs cannot do language for 50 years.



Hinton: analogy with another paradigm shift: continental drift. Geologists were confident Earth was rigid. Holmes suggested a plausible mechanism but was ignored.

Hinton: if you train a model with hundreds of billions of parameters

Central question: Do LLMs understand what they are saying?

Turing test, Winograd sentences

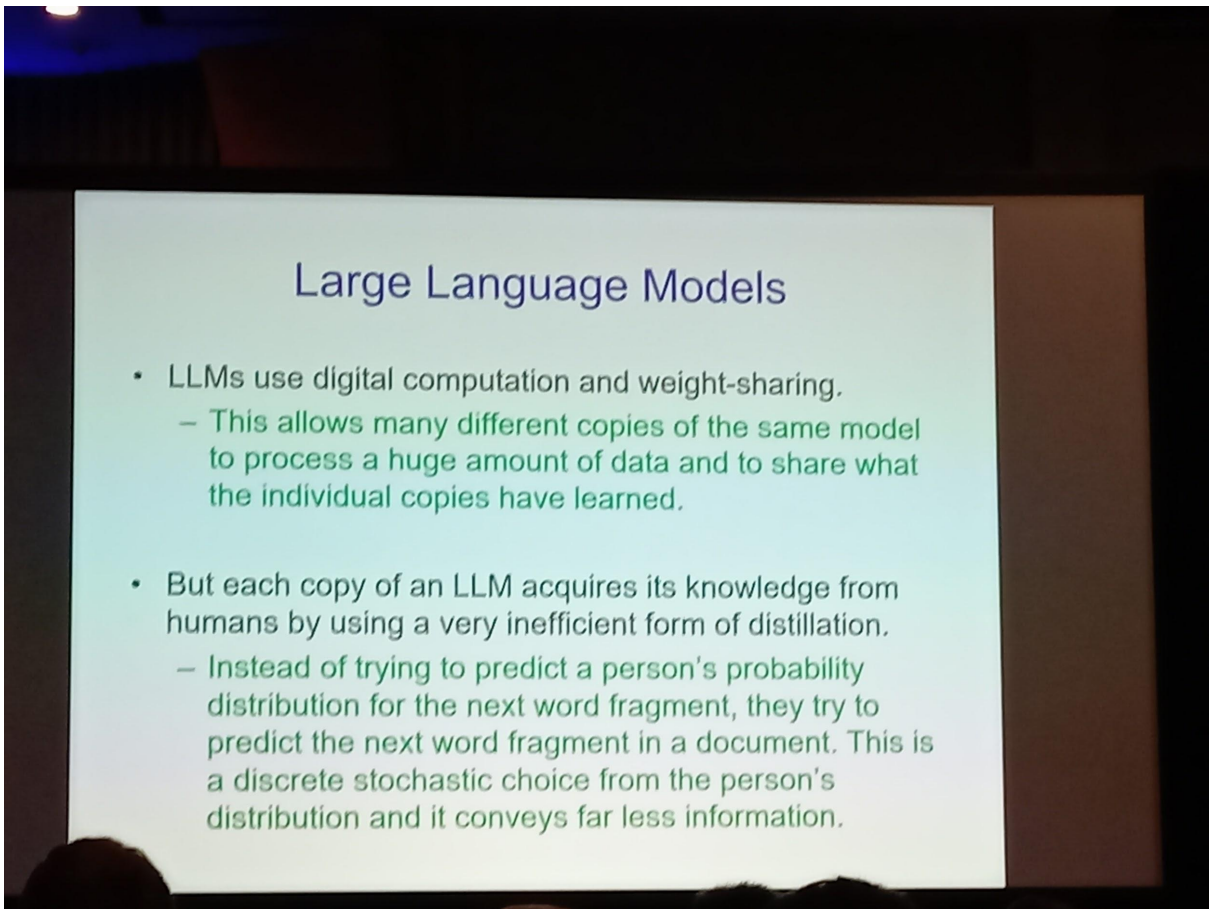
Hinton: I don't see how GPT-4 can answer how it does without understanding the questions. Of course, it does not understand everything.

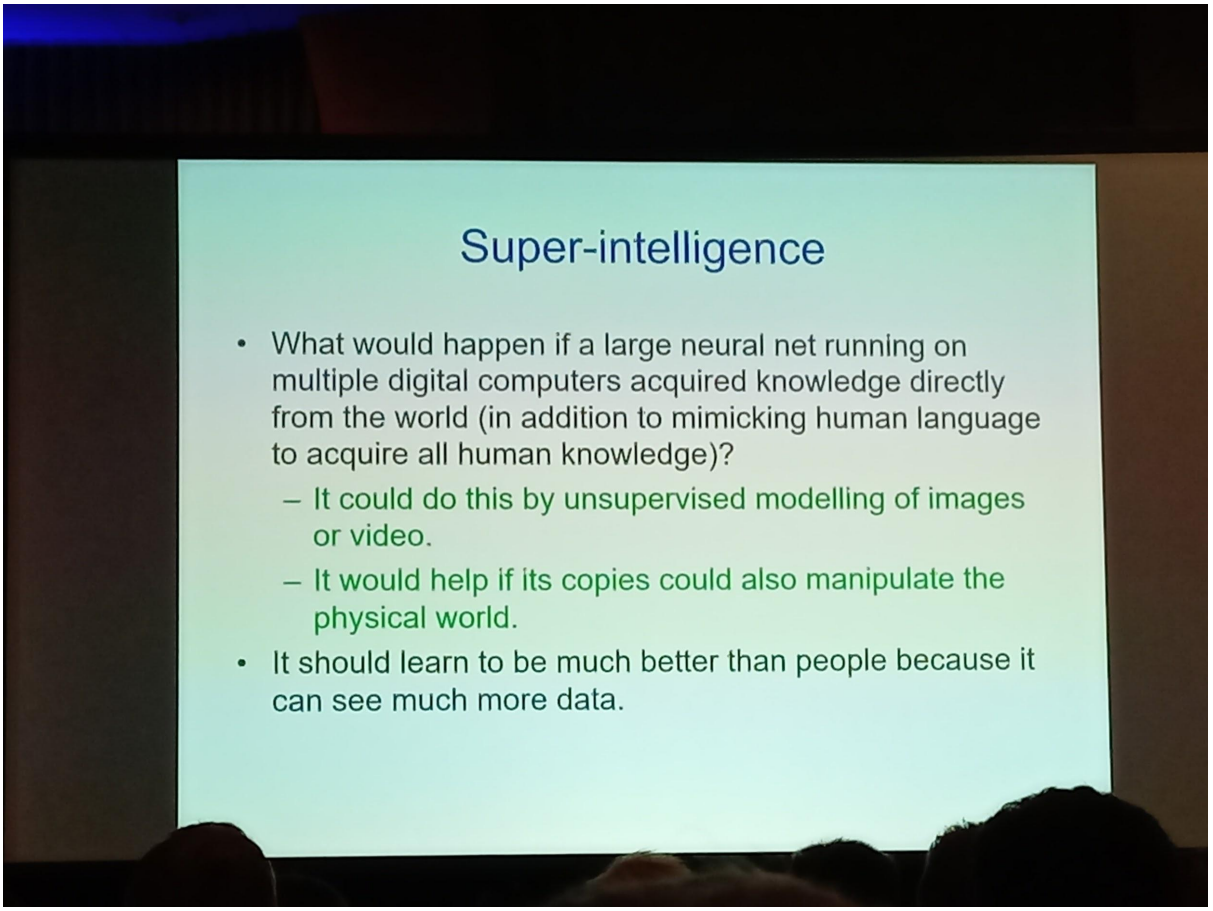
"They should not be called hallucinations, they should be called confabulations."

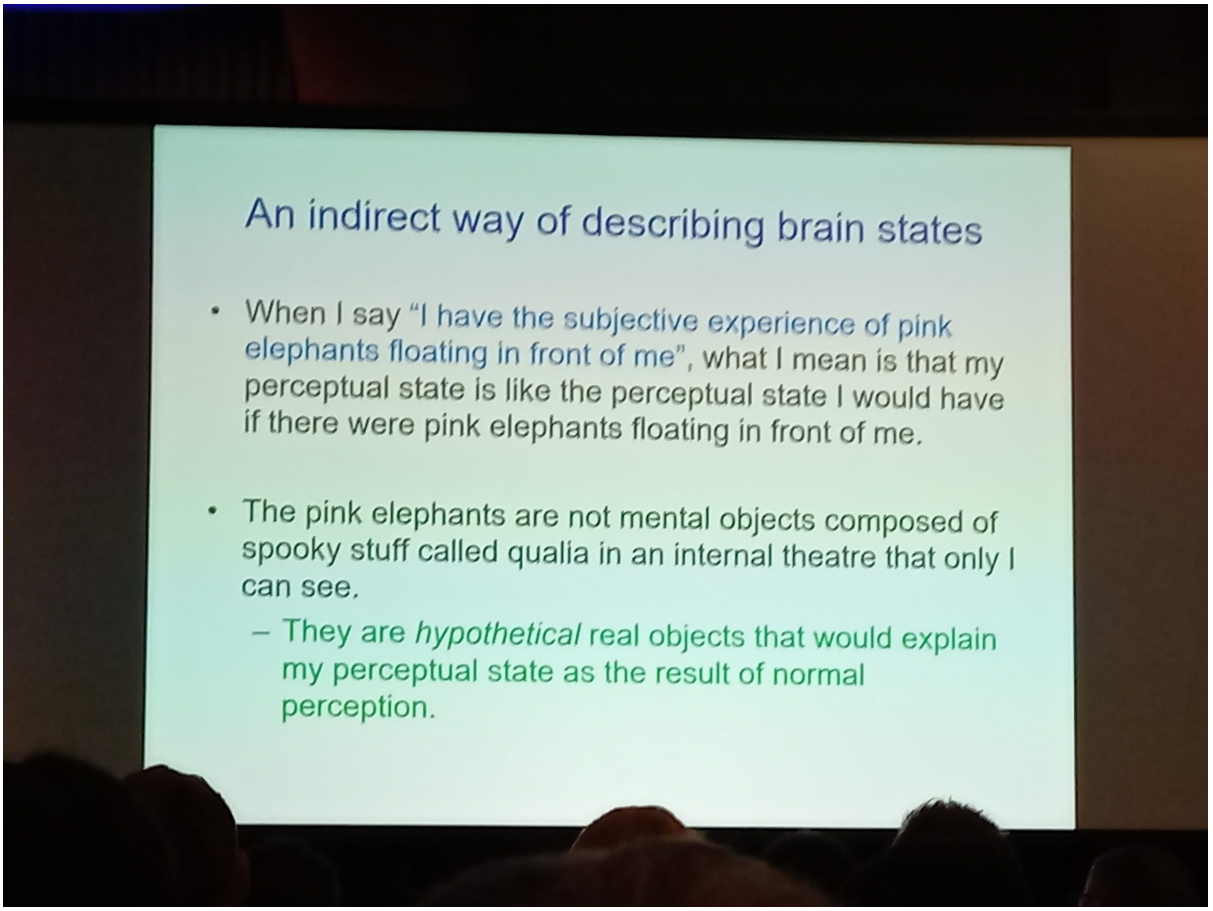
"LLMs confabulate. People confabulate all the time."

But isn't it just predicting the next using statistical word co-occurences? It's just statistics.

Hinton: "No, that's not what's happening. It's not just statistics."


Large Language Models
• LLMs use digital computation and weight-sharing.
– This allows many different copies of the same model to process a huge amount of data and to share what the individual copies have learned.
• But each copy of an LLM acquires its knowledge from humans by using a very inefficient form of distillation.
– Instead of trying to predict a person's probability distribution for the next word fragment, they try to predict the next word fragment in a document. This is a discrete stochastic choice from the person's distribution and it conveys far less information.


Super-intelligence
• What would happen if a large neural net running on multiple digital computers acquired knowledge directly from the world (in addition to mimicking human language to acquire all human knowledge)?
– It could do this by unsupervised modelling of images or video.
– It would help if its copies could also manipulate the physical world.
• It should learn to be much better than people because it can see much more data.



Emily Bender is the first to ask a question :-)))

Emily: What do you mean if you say a computer understands a question and how do you test that in a robust way?

Vivid discussion between Emily Bender and Hinton mostly revolves around definition of understanding.

The queue for questions is long. People jump from their chairs, talk vividly, clap their hands or disagree or shake their heads.

The queue now reaches third of the aisle. Both left and right aisles.

Cannot capture the exact arguments, it's a storm.

Last question announced! Too bad, there's still so many people in the queue.

Random guy leaving the ballroom (about Hinton): "He takes such a strong position."

Personally, I don't take as strong position as both camps, but Hinton's "computers understand because I tricked someone to think a text generated by GPT-4 was my text" and "computers describe things and therefore have subjective experience" is too strong. Also, the talk lacked good arguments. "I don't see how they could not understand" is not a

scientific argument. Finally, engineering is not everything. Does he read philosophy? Psychology? Psycholinguistics?

Monday Session 1: Poster Session

See poster photos

Monday Session 2 Reality Checks

Credible Without Credit

Are ChatGPT and YouChat reliable? They asked questions and had experts rate the answers.

0/10 experts recommended YouChat for professional use. 3/10 experts recommended ChatGPT for professional use.

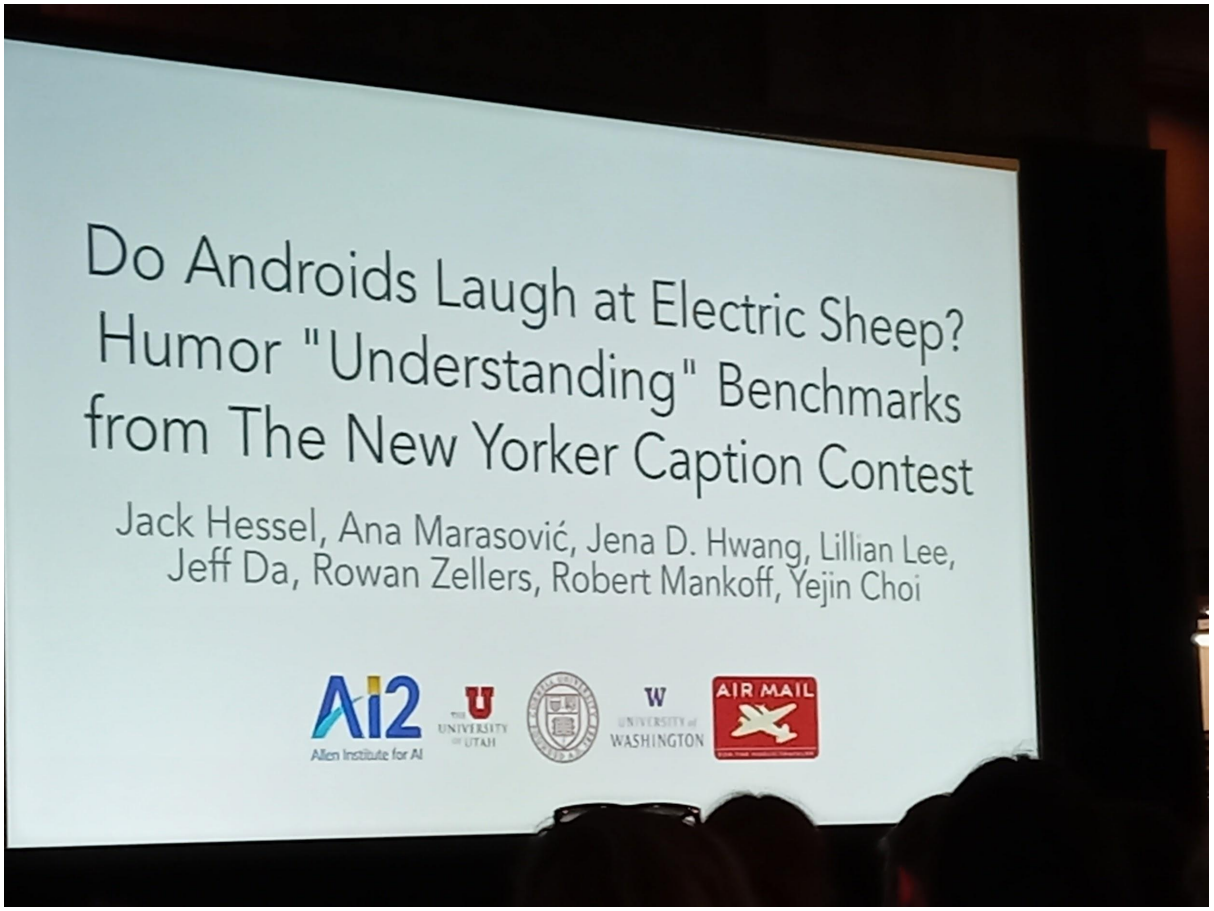
What’s the Meaning of Superhuman Performance in Today’s NLU?

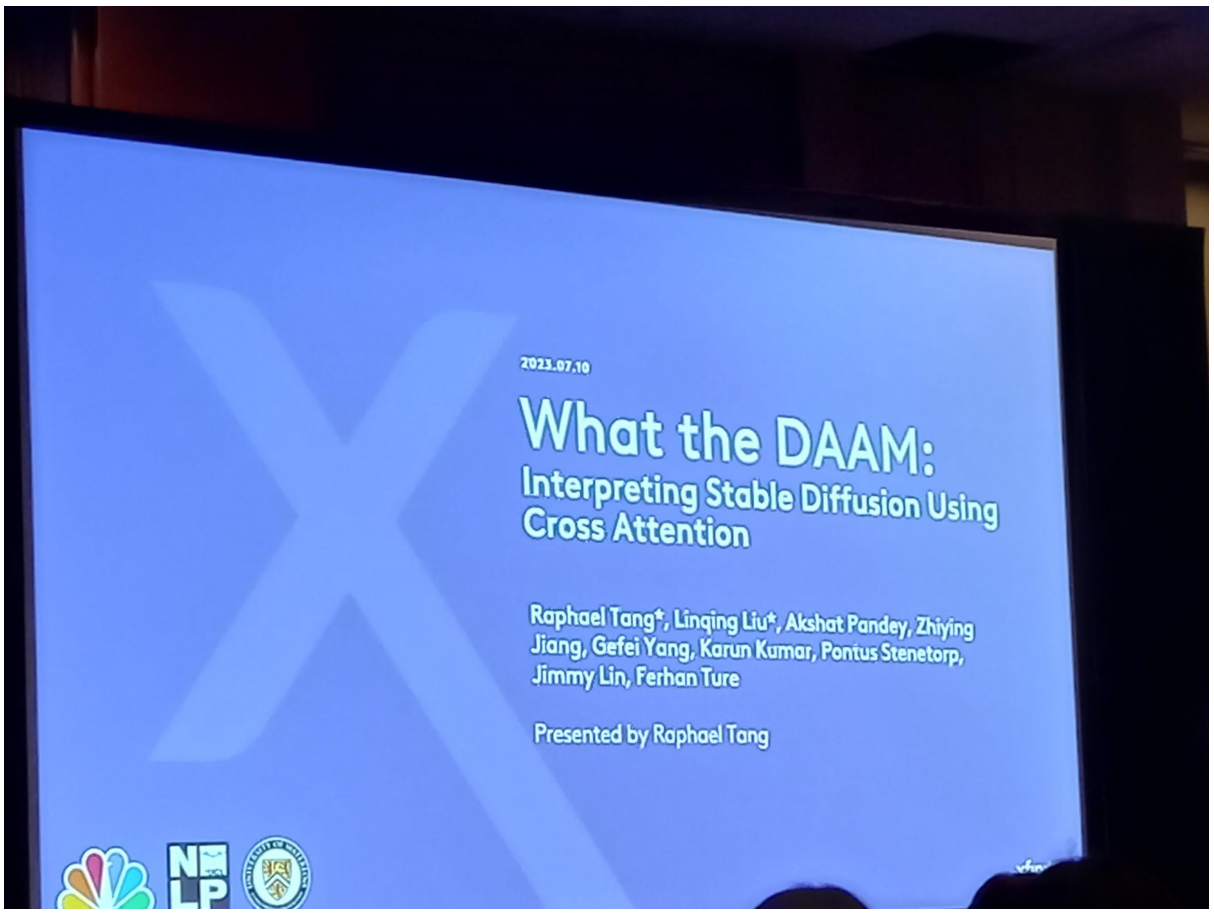
Really good paper which deserves the Outstanding Paper Award. Great presentation too.

Why Aren't We Ner Yet?

Robustness of NER in ASR.

Best Paper Awards




2023.07.10
What the DAAM:
Interpreting Stable Diffusion Using Cross Attention
Raphael Tang*, Linqing Liu*, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, Ferhan Ture
Presented by Raphael Tang
NLP

Tue Session 3 LLMs

Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages (Area Chair Award)

Supports 500 languages (XLM-R 100).

Tail languages. Not covered by XLM-R. Little data available.

Training data: Found 2000 languages, kept only those with > 30k sentences.

Vocabulary extension: 250k + 150k = 400k

In many metrics, performs better than XLM-R.

mCLIP: Multilingual CLIP via Cross-lingual Transfer

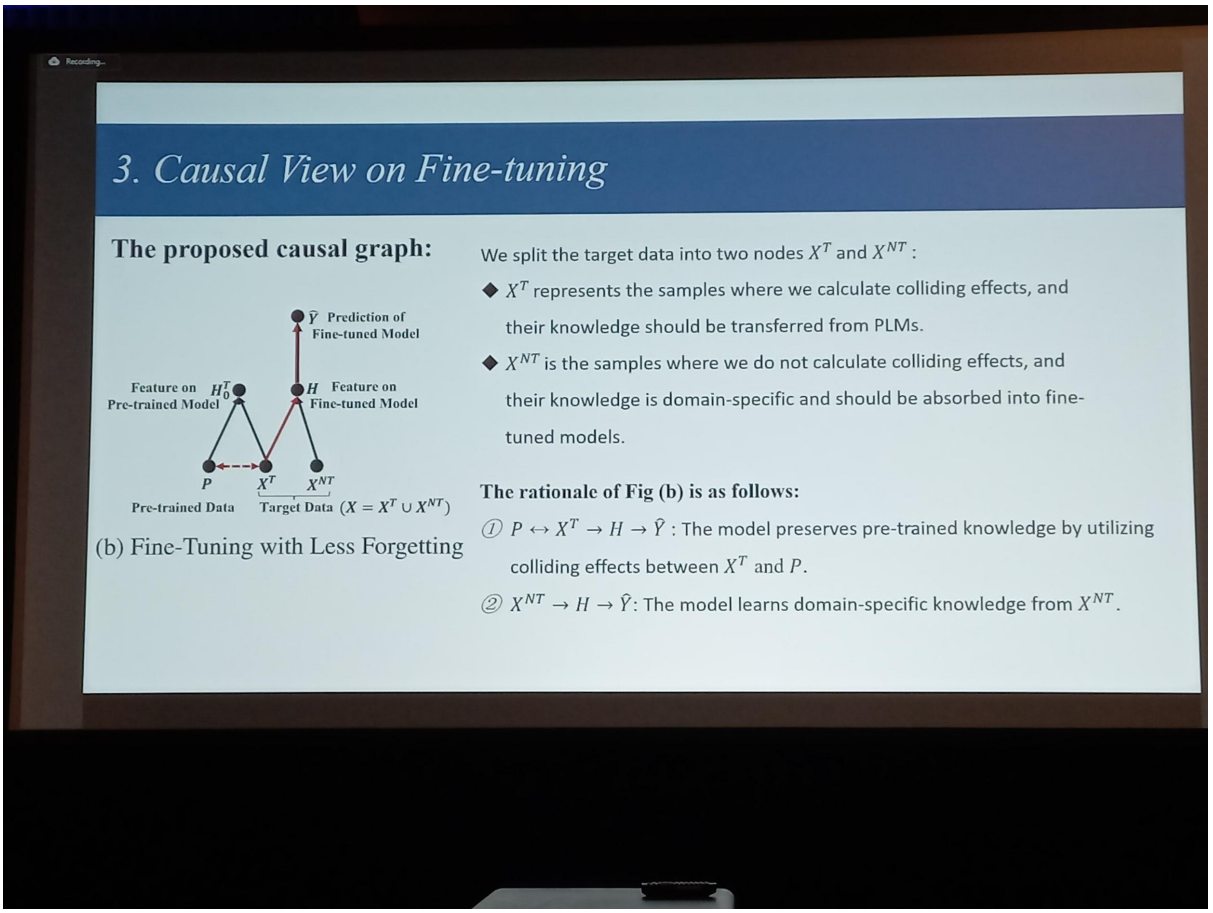
Multilingual CLIP. The talk was streamed, with lots of noise and technical details so I got lost. The presentation was not easy to follow. But I think the work is very good. Adds multilinguality and also improves over CLIP.

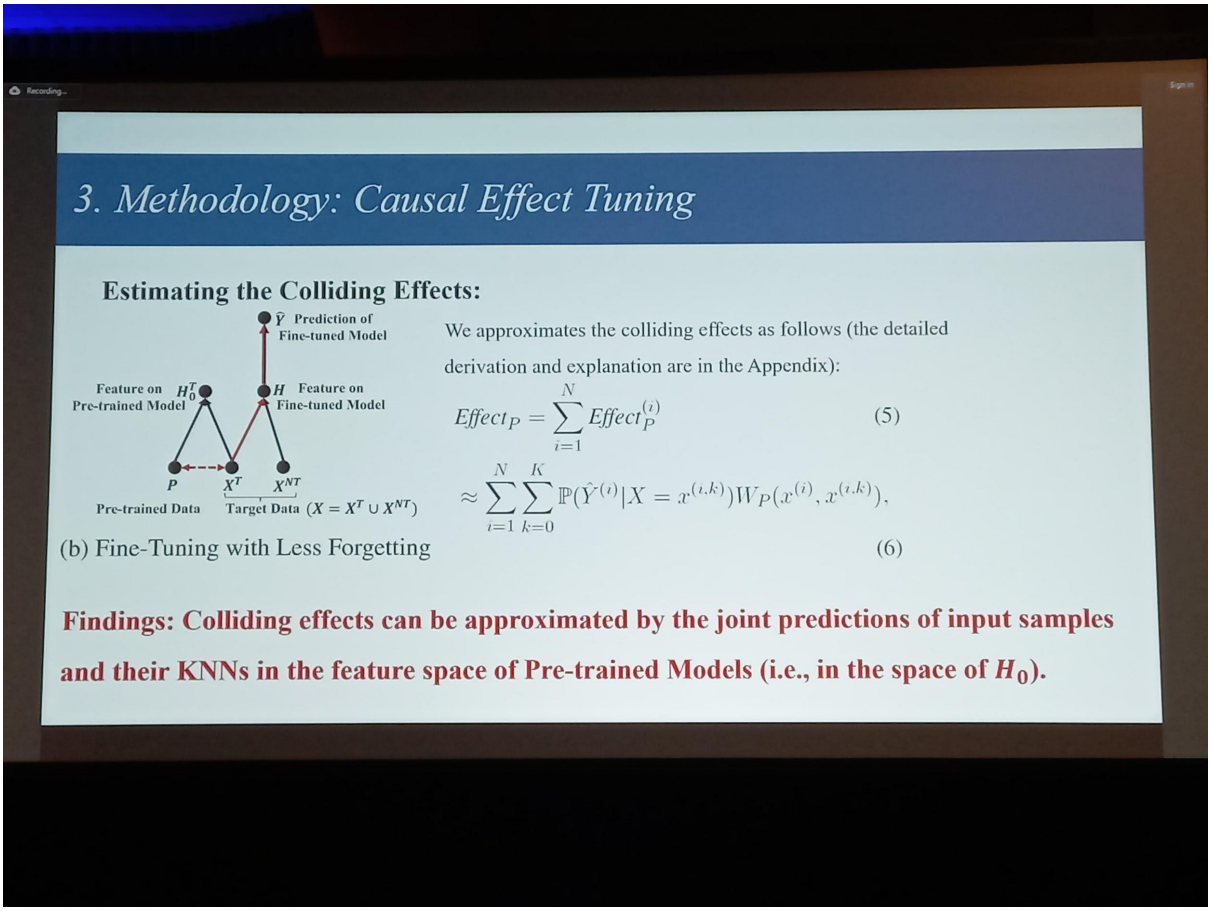
Preserving Commonsense Knowledge from Pre-trained Language Models via Causal Inference

How to prevent catastrophic forgetting of the pre-trained information during fine-tuning.

Vanilla fine-tuning forgets data from pre-training, because it is missing the original pre-training data.

Fine-tuning with less forgetting:





Zlepšili SOTA na spoustě QA úloh. Přitom nepotřebovali KG (knowledge graph). Ale když ho ještě přidali, tak to ještě zase zlepšili.

Taky pěkná čísla na NERu (92 něco se starým fine-tuningem a 92 víc s novým).

Self-Adaptive In-Context Learning: An Information Compression Perspective for In-Context Example Selection and Ordering

Incorporates MDL (minimum description length) into loss. The paper brings interesting insight.

Pre-Training to Learn in Context

From pre-training to ICL (in context learning). Discrepancy between pre-training and the ICL format.

(ICL = in-context learning = fancy name for prompting, e.g. give a few examples with labels as part of the prompt and then adk for the label of a new example)

Approach: "intrinsic task". Large-scale plain text corpus naturally contains intrinsic tasks (e. g. paragraph ending task. Not NER - classes missing).

Train on downstream datasets, generalization to unseen (intrinsic) tasks.

(I wrote it down but the talk was fast and hard to understand due to accent so I don't really know what it was about.)

RegAugKD: Retrieval-Augmented Knowledge Distillation For Pre-trained Language Models

NObvyklé KD si po provedení nezachovává žádnou informaci od teacher modelu a ten se zahodí a už se při inferenci student modelem nepoužije.

Tady si uloží neparametrickou paměť příkladů od učitele.

Konstrukce paměti: na učitele se naloží jednoduchá poslední vrstva (netřeba přetrénovat učitele, jen tu vrstvu).

Trénuje se s novou loss, která minimalizuje vzdálenost mezi teacher-teacher a teacher-student embeddingy.

Pak se konstruuje knowledge base obsahující projekci teacher embeddingů a predikce.

Při inferenci se použije kNN do téhle knowledge base k teacher embeddingům a z nich se použijí predictions.

Výstup je naučený vážený průměr z predikcí učitele a studenta.

Tue Session 4 LLMs

RetroMAE-2: Duplex Masked Auto-Encoder for Pre-training Retrieval-Oriented Language Models

Špatná prezentace, nesrozumitelná.

Augmentation-Adapted Retriever Improves Generalization of Language Models as Generic Plug-In

Taky se nedalo poslouchat.

Pre-trained Language Models Can be Fully Zero-Shot Learners

Chytré a pěkné. NPPrompt.

Summary

- **NPPrompt**, a novel method for **fully zero-shot** learning with pre-trained language models (PLMs).
- NPPrompt utilizes PLMs' **initial word embeddings** to identify related words for category names, without manual design or unlabeled data.
- Empirical results show that NPPrompt significantly outperforms previous fully zero-shot methods.

Takeaway

NPPrompt can be easily plugged into any SOTA LLM

- Employ k-Nearest-Neighbor in LLM's token embedding space
- Nonparametric aggregation
- **Efficient natural language understanding**
- Dynamic zero-shot problems

Thanks for your listening!

Homepage: https://xuandongzhao.github.io/
Email: xuandongzhao@ucsb.edu

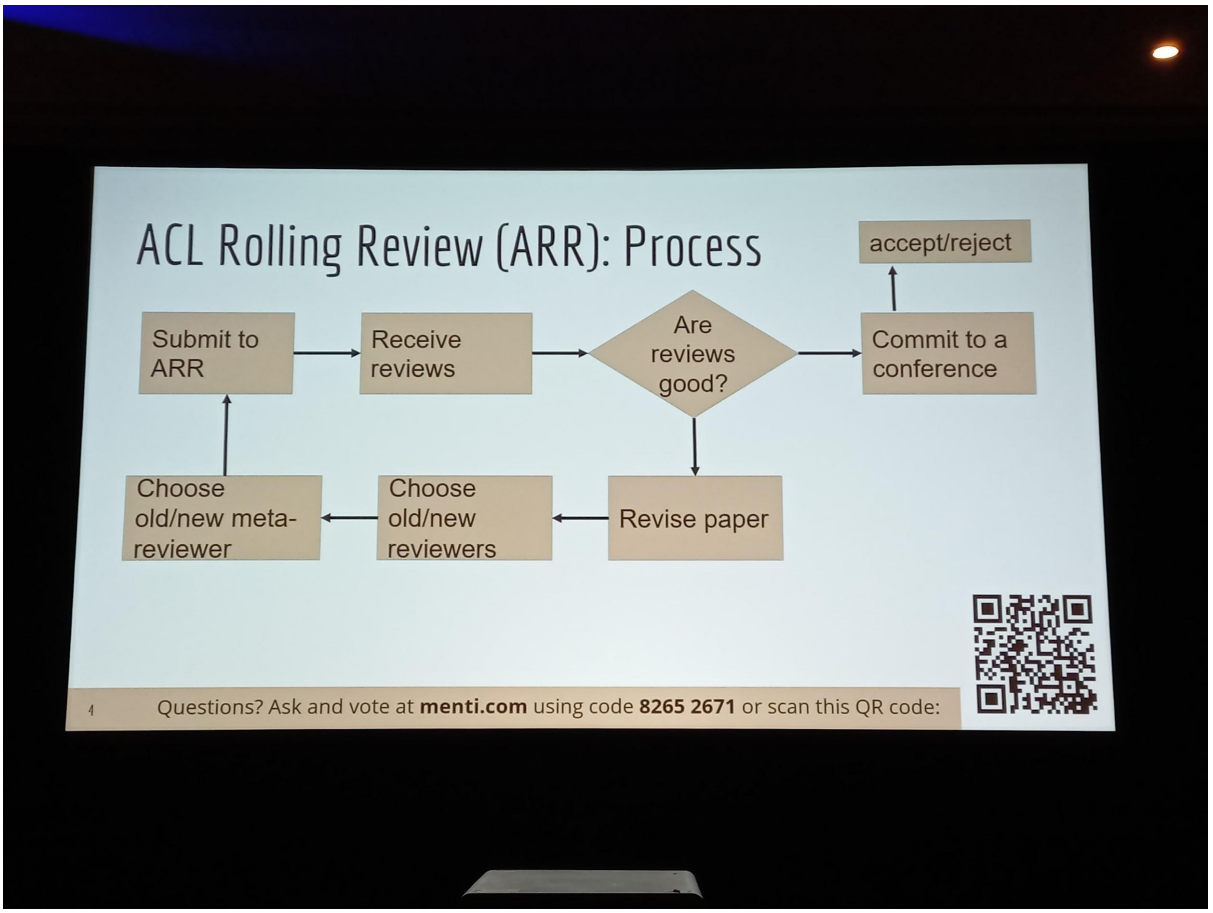
SCAN ME

Multilingual LLMs are Better Cross-lingual In-context Learners with Alignment

Surface-Based Retrieval Reduces Perplexity of Retrieval-Augmented Language Models

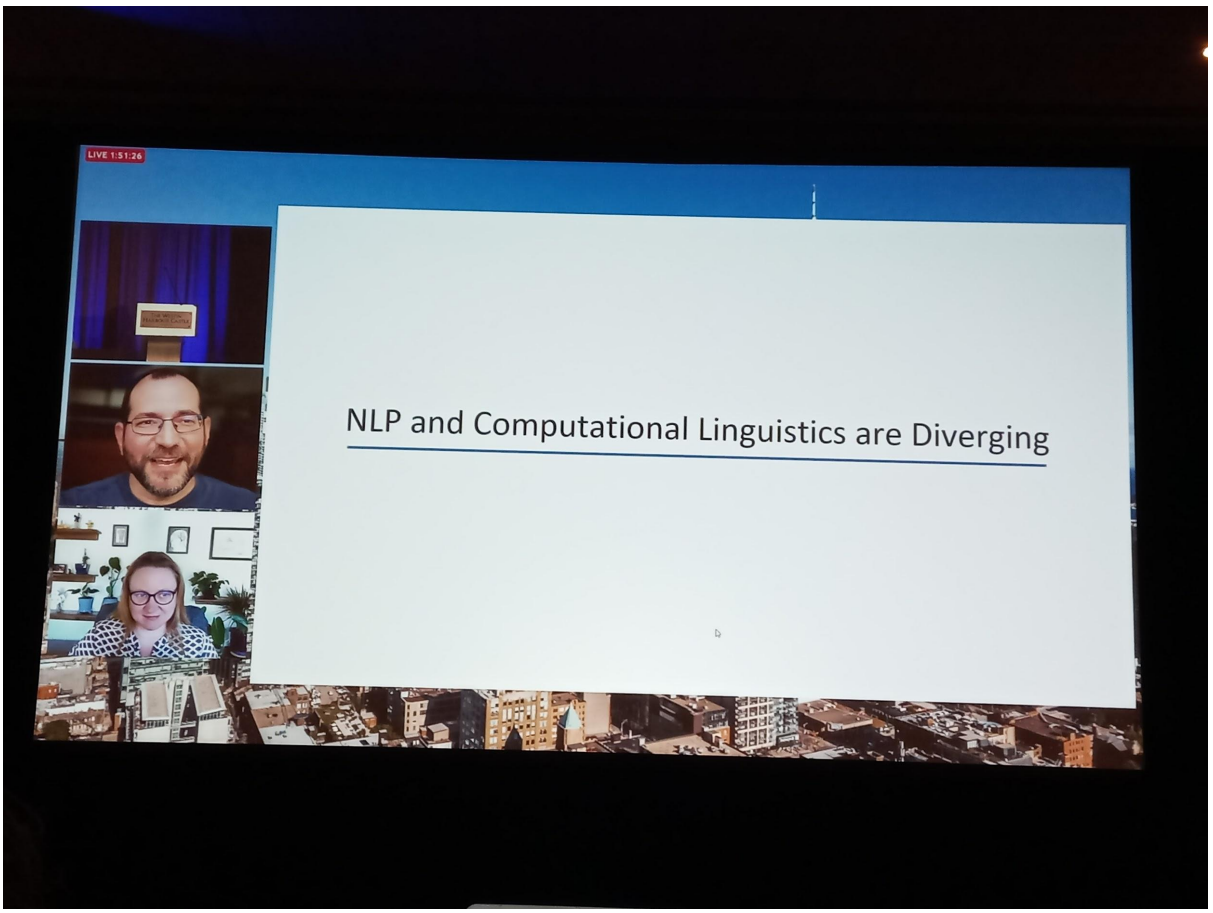
Understanding In-Context Learning via Supportive Pretraining Data

ACL Rolling Review



Panel:

Dan Klein:



- NLP does not solve computational linguistics any more
- Scientific processes yielding to commercial processes
- NLP systems going from too bad to too good

Margaret Mitchell:

Linguistic Theories, Cognitive Modeling and Psycholinguistics

Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times?

Expectation-based theories of sentence processing Hale 2001, Levy 2008

Surprisal effects

LM perplexity vs. human reading times

GPT-2 positive relationship

Evaluate surprisal to predict reading times.

Bigger LMs get worse at predicting.

Snaží se to vysvětlit lingvistickými jevy.

Underpredicting, overpredicting.

They conclude we should use smaller models.

My notes: NOOO never they use the probability theory wrong. Must write reply to this paper!

Session Machine Translation

Interesting works on multilingual NMT (one model for all language pairs).

Tuning the sampling temperature is key for performance.

Kvůli focení jsem skončila na Machine Translation Session, ale není to špatné. Hodně se zabývají multilingual NMT, což je jeden megamodel pro všechny jazykové páry. A popisují různé strategie a problémy.

Zajímavé je, že všichni se zabývají tím, jak samplovat jazyky do trénovacích dat a používají vždy temperature sampling.

Říkají, že pohrát si s hyperparanetrem teploty je důležité pro výsledný performance.

Nám to tak v shared tasku loni na CorefUD nevyšlo, bylo skoro jedno, jak jsme data z datasetů mixovali.

ALE: tohle se týká překladových modelů a ne fine-tuningu na downstream tasks.

Pak se taky lidi hodně zabývají evaluací MT (slabiny COMET, …)

Poster Session

Do CoNLL-2003 Named Entity Taggers Still Work Well in 2023? (Ouststanding paper)

Keynote: Alison Gopnik

The typical way we think about LLMs is really misguided.

The agent view on AI. Individual agents. Are they super intelligent? Are they evil?

The picture is that AI systems always lead to disaster.

We think the way to evaluate agents or person or system is in terms of intelligence.

This is deeply misguided.

It is not the right way of thinking, of evaluating.

What is then?

Large models as cultural technology.

Such as language, writing, print, library, internet, wikipedia.

Cultural technology is where we get information. We can use it to our advantage. Language: information technology built into our biology.

Cultural evolution: imitation vs innovation. Cognitive capacities for imitation. Cognitive capacities for innovation allow exploration, new discoveries, causal discovery, induction.

These concepts are different from each other.

The hypothesis is that LLMs will be really good at imitation, not at discovery.

Which cognitive capacities are enabled of facilitated by a new cultural technology?

There's no such thing as general intelligence - natural or artificial.

Developmental psychology does not describe general intelligence. It's just incoherent from cognitive science point of view. Developmental cognitive science describes a vast array of capacities.

Closing Ceremony

Lifetime award: Martha Palmer