CS 333: NLP

Fall 2023

Prof. Carolyn Anderson
Wellesley College

PLEASE JOIN US FOR A

CS Holiday Party

ALL CS, MAS, DS AND CLCS MAJORS AND MINORS.

THERE WILL BE TEA, SNACKS, GAMES AND GINGERBREAD HOUSES!

THURSDAY, DECEMBER 14, 2023
TIME: 4-6 PM
SCIENCE CENTER LEAKY BEAKER

RSVP Here

Questions:
cdelcour@wellesley.edu

Disability Accommodations:
disabilityservices@wellesley.edu
or 781.283.2434

WELLESLEY

Help hours

My help hours for the rest of the semester:

- Today: 3:30-4:30
- Monday: 3:30-4:30
- Thursday: 3-4

Final Project Presentations

- 2-3 minutes per person
- 1 slide (upload to Google Drive by 7am on Tuesday)

What Have We Learned?

Tasks We Learned This Semester

- Tokenization
- Sentence Segmentation
- Textual Similiarity Analysis
- Representation Learning
- Classification: sentiment analysis, genre prediction...
- Language Modeling
- Common-sense reasoning

Techniques We Learned This Semester

- Tokenization: BPE
- Textual Similiarity Analysis: tf-idf, PPMI
- Representation Learning: term-document frequency matrixes, static word embeddings, contextualized word embeddings
- Classification: Naive Bayes, regression, neural networks
- Language Modeling: n-grams, neural networks (feedforward, recurrent, attention-based)
- Transfer Learning: finetuning, prompt engineering

Models We Used This Semester

- word2vec
- DistilBERT
- feedforward neural networks in PyTorch
- LLaMA

Skills We Developed This Semester

- text processing: working with strings, reading in text datasets, tokenizing, cleaning text, regular expressions
- evaluation: computing metrics (accuracy, precision, recall, F-score, perplexity, and cosine similarity), understanding statistical reliability
- implementing classification models
- building and running neural network models
- finetuning and prompting large language models

What Is There Still To Learn?

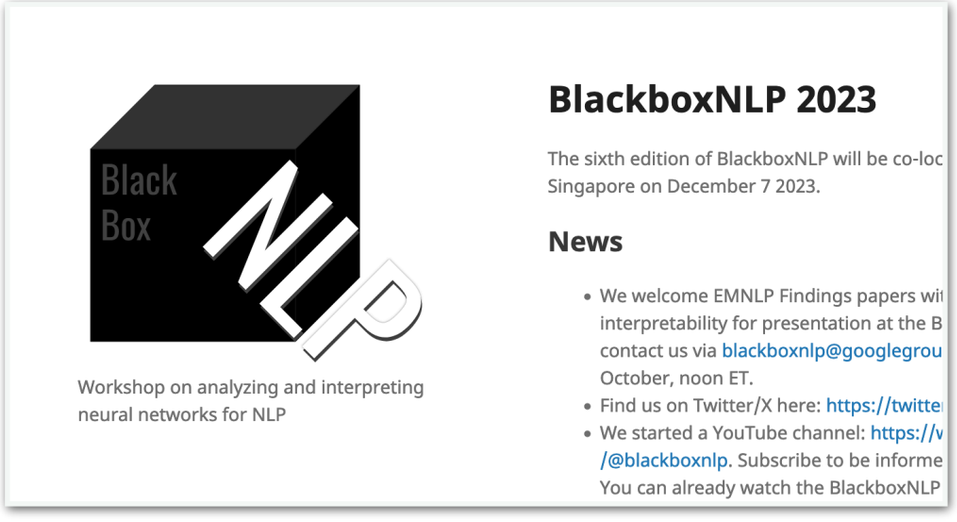
Interpretability and Explainability

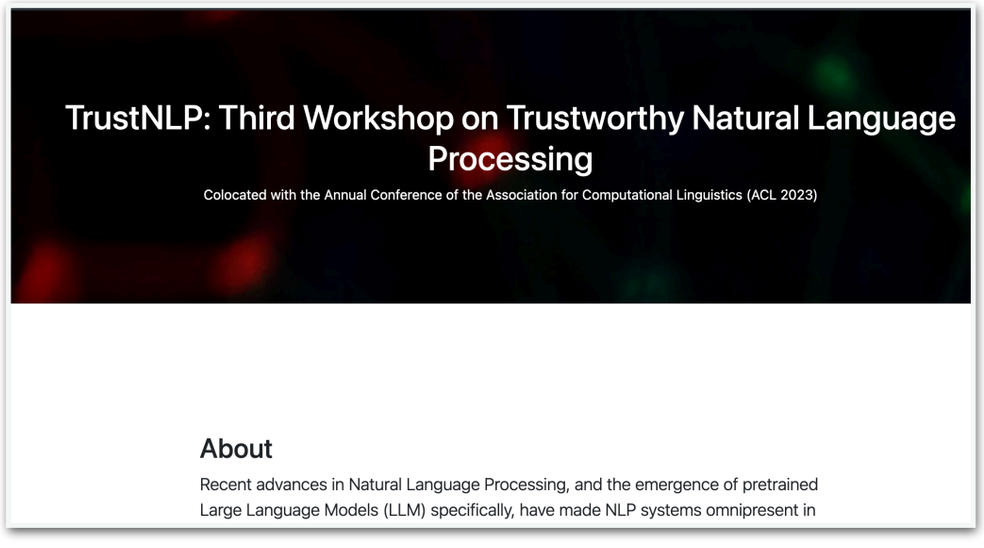
How do blackbox models make decisions?

What are large language models learning, and how do they use this information?

What are the limits of what large language models?

How can we intervene when models are incorrect?





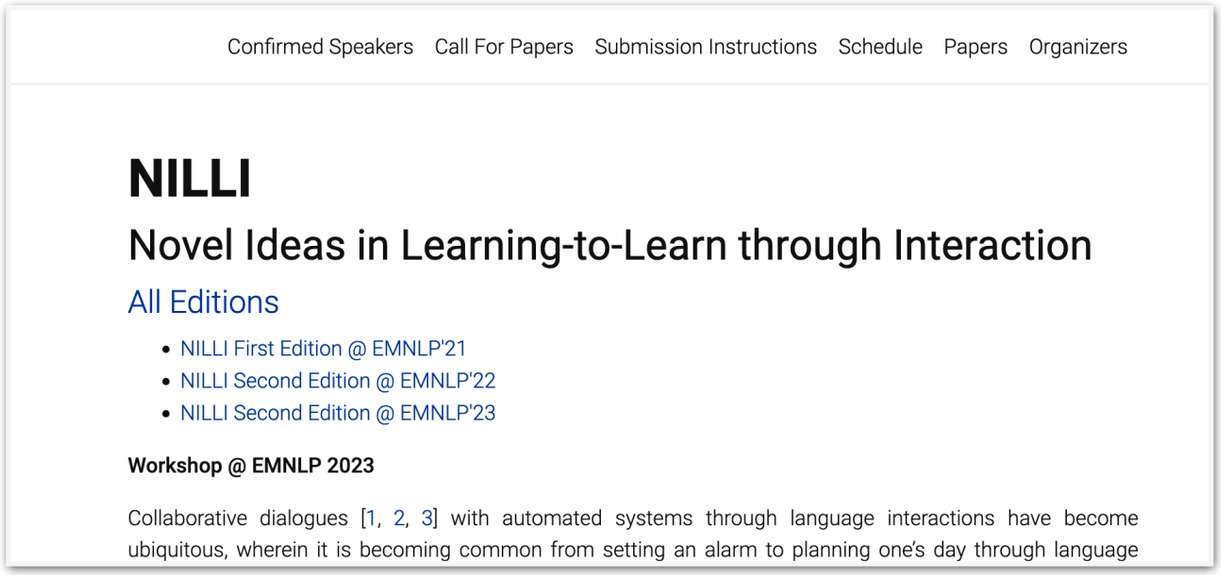
Grounding and Multi-Modality

How do we combine text-based models with other modalities, like vision and perceptual information?

How can we help models relate to the physical world?

How do we help models understand social interaction?

How much of a role does access to perceptual information play in learning?



Linguistic Diversity

How can we improve NLP tools for low-resource languages?

How can we achieve good performance with sparse data? With small models and limited compute?

How do models deal with different language varieties?

Computational Social Science

How can we share our best approaches for text processing with other fields?

How well do NLP techniques work with limited data? In niche domains? In tasks that require expert judgment?