CS329X: Human Centered NLP

Human-Centered NLP

Diyi Yang
Stanford CS
Welcome to CS 329X
Instructor and CA

Diyi Yang

Rishi Bommasani
Course Overview

Website:
http://web.stanford.edu/class/cs329x

Ed Discussion:
https://edstem.org/us/courses/38475/discussion/
What is human-centered NLP?

“What human-centered NLP involves designing and developing NLP systems in a way that is attuned to the needs and preferences of human users, and that considers the ethical and social implications of these systems.”

– ChatGPT, 2022
What is human-centered NLP?

It concerns NLP systems, which goes beyond just the model – also includes e.g. user interfaces on top of the model.

“Human-centered NLP involves designing and developing NLP systems in a way that is attuned to the needs and preferences of human users, and that considers the ethical and social implications of these systems.”

– ChatGPT, 2022

It needs to be optimized for humans.

“Optimize for humans” require careful ethical concerns.

It touches multiple NLP dev stages.
What if this definition is wrong?

What is a right definition of human-centered NLP?
Why should we build human-centered NLP?

“The common misconception [is] that language use has primarily to do with words and what they mean. It doesn’t. It has primarily to do with people and what they mean.

Corrective
Preventive
Not Reactive
Who is the “human” in human-centered NLP?

The "human" refers to the focus on designing and developing NLP technologies that prioritize human needs and preferences, rather than solely focusing on technological capabilities.

Human-centered NLP seeks to create NLP systems that are more user-friendly, accessible, and inclusive.
Human-centered NLP should be in every stage

1. Task Formulation
2. Data Collection
3. Data Processing
4. Model Training
5. Model Evaluation
6. Deployment
Human-centered NLP is a spectrum
Traditional Example on Data Collection

"Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text."

See passage

Generate Question

Traditional Example on Data Collection: Issues

Crowdworkers are hired to generate questions in a constrained setting, which could be different from how people generally ask questions. Crowdworkers generate questions specific to one paragraph & are primed to generate questions of a certain style. Questions created in this way have a high degree of lexical overlap with the document text and thus models might rely too heavily on word matching.

Q: What component of water is more soluble than nitrogen?

Doc: Oxygen is more soluble in water than nitrogen is. Water in equilibrium with air contains approximately 1 molecule of dissolved O2 for every 2 molecules of N2, compared to an atmospheric ratio of approximately 1:4. ...

A: Oxygen

Q: What is the world's largest academic and private library system?

Doc: Harvard ... operates several arts, cultural, and scientific museums, alongside the Harvard Library, which is the world's largest academic and private library system, comprising 79 individual libraries with over 18 million volumes...

A: Harvard Library
**Google Natural Question:** Questions consist of real anonymized, aggregated queries issued to the Google search engine. An annotator is presented with a question along with a Wikipedia page from the top 5 search results, and annotates a long answer (typically a paragraph) and a short answer (one or more entities) if present on the page, or marks null if no long/short answer is present.
More Human-Centered Data Collection

Data collection schema, if designed unnaturally, can introduce annotation artifact (unwanted patterns in the data), and can be harmful or unfair to certain groups.

Example: Image classifier trained on the Open Images dataset that does not properly apply “wedding” related labels to images of wedding traditions from different parts of the world.

Human-Centered data collection should focus on mimicking real-use scenarios so the data will reflect actual human needs.
Traditional Model Training

“The GPT-3 language model can be coaxed to perform natural language tasks using carefully engineered text prompts. But these models can also generate outputs that are untruthful or toxic. This is in part because GPT-3 is trained to predict the next word on a large dataset of Internet text, rather than to perform the language task that the user wants.”

<table>
<thead>
<tr>
<th>PROMPT</th>
<th>COMPLETION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explain the moon landing to a 6 year old in a few sentences.</td>
<td>GPT-3 Explain the theory of gravity to a 6 year old. Explain the theory of relativity to a 6 year old in a few sentences. Explain the big bang theory to a 6 year old. Explain evolution to a 6 year old.</td>
</tr>
</tbody>
</table>
Maybe More Human-Centered Model Training

PROMPT

Explain the moon landing to a 6 year old in a few sentences.

COMPLETION

GPT-3

Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

InstructGPT

People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.

Reinforcement learning from human feedback (RLHF): make models safer, more helpful, and more aligned

“On prompts submitted by our customers to the API, our labelers provide demonstrations of the desired model behavior, and rank several outputs from our models. We then use this data to fine-tune GPT-3.”
“The resulting InstructGPT models are much better at following instructions than GPT-3… Our labelers prefer outputs from our 1.3B InstructGPT model over outputs from a 175B GPT-3 model, despite having more than 100x fewer parameters.”

“it “unlocks” capabilities that GPT-3 already had, but were difficult to elicit through prompt engineering alone.”

Compared to standard training (which gives models many hidden power), slightly more human-centered training might make models more usable.
Traditional Evaluation

“performance on the benchmark has recently come close to the level of non-expert humans, suggesting limited headroom for further research.”

Accuracy on benchmarks is the most standard assessment on model quality, but it *does not* contain enough signal!

Traditional Evaluation

Shortcuts/right for wrong reasons

What is the moustache made of?
> Banana

What are the eyes made of?

What is?
> Banana

What?
> Banana

More Human-Centered Evaluation

“Capability testing”, quantify human expectations and requirements on models

<table>
<thead>
<tr>
<th>Test name</th>
<th>failure rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>change neutral words with BERT</td>
<td>51 / 500 = 10.2%</td>
</tr>
</tbody>
</table>

Test Summary

Test [INV] on [VOCABULARY]
change neutral words with BERT

Desc. Change a set of neutral words with other context-appropriate neutral words (using BERT).

Result FAILURE RATE ON ALL CASES
51/500=10.2%

Examples

- @united I will not be flying you back again. Pred: 2 (0.71) -- 0 (0.99)
- @AmericanAir you-- never called back just to put me on hold. It's midnight. Literally just want to know how I'm getting home and I'm getting no help. Pred: 2 (0.81) -- 0 (0.99)
- @AmericanAir Yes I am. 2495/1 170. RNO departure at 1229 on 2/25 w/connection at DFW to-- and LGA. I can do the

More Human-Centered Evaluation

Standard evaluations cannot capture model shortcuts.

Instead, human-centered evaluation should design fine-grained metrics and analysis strategies that account for user-specific interaction objectives, cognitive loads, etc.
Some parts of a task is inherently hard. For example, nouns in some language is gender-specific but not in others, and translator won’t be able to “add” this information, so the disambiguation needs to be handled by the interface.
Human-Centered Deployment

Similar models also need to have very different interaction wrappers in different use cases.

Human-Centered Deployment

Who is going to use the system?
Who is going to design the system?
How would users use the system?
What interface can best facilitate such interaction?

Handling exceptions & designing interactions for the right user group goes beyond modeling, and is essential for deployment.
Why should we care about human-centered NLP

Improve user experiences

Build effective, accountable and responsible systems

Benefit the society as a whole
What if our NLP systems are not human-centered?

- Biased results
- Inaccurate interpretation
- Lack of culture awareness
- Lack of personalization
- Security risks
- …
AI that’s not human-centered can be harmful

Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process.

When Bias Is Coded Into Our Technology

February 8, 2020 • 6:03 AM ET

By Jennifer B. Lee

Amazon’s automated hiring tool was found to be inadequate after penalizing the résumés of female candidates. Photograph: Brian Snyder/Reuters

Courtesy of the 2020 Group
AI that’s not human-centered can be harmful

There Is a Racial Divide in Speech-Recognition Systems, Researchers Say
Technology from Amazon, Apple, Google, IBM and Microsoft misidentified 35 percent of words from people who were black. White people fared much better.

NEWS & COMMENTARY

Algorithms Are Making Decisions About Health Care, Which May Only Worsen Medical Racism

Unclear regulation and a lack of transparency increase the risk that AI and algorithmic tools that exacerbate racial biases will be used in medical settings.
Modern AI Systems are learning from human preferences

https://openai.com/blog/chatgpt
How to Make NLP More Human-Centered?

**NLP people** know the standard method of data preparation, training, evaluation, and deployment.

**HCI people** know ways to mimic natural use scenario, collect human feedback, design interactions…

Both are needed for Human-Centered NLP.
What do you want to learn from CS329X?

Join at

slido.com

#3615 081
Course Goals

The primary goal of the course is **offer an overview of HCI+NLP**, and to help students get access to, and understand, both HCI and NLP research papers and methods. More specifically, we will:

- Introduce basic concepts of NLP and HCI
- Introduce a variety of emerging topics related to human-centered NLP

**Work together to think about and define what is human-centered NLP**

This course is co-designed with Prof. Sherry Wu at CMU.
Course Topics

**NLP basics**
1. The ultimate crash: NLP tasks and applications
2. Human-centered design principles
3. Human in the loop
4. Learning from human feedback
5. Human-centered evaluation
6. Data collection and curation
7. Beyond benchmarking
8. Interpretability and explanation
9. Privacy, security and safety
10. Trust in AI
11. Human-AI collaboration
12. NLP for social impact

**Model Design**

**Data and Evaluation**

**Safety/Trust**

**Impact**

NLP basics

Model Design

Data and Evaluation

Safety/Trust

Impact
Guest Lectures on Emerging Topics

Sherry Wu (CMU)  
Interactive NLP and Visualization

Rishi Bommasani (Stanford)  
Ethics and Social Responsibility

Joon Sung Park (Stanford)  
Simulation via Foundation Models

Irene Solaiman (Huggingface)  
AI Governance
What does CS 329X provide?

**Provides** an overview of HCI + NLP

**Provides** various aspects of what a usable NLP system looks like

**Does not provide** an in-depth intro to deep learning for NLP (see CS224N)
Prerequisites

CS 224N or CS224U, or equivalent.

You are expected to…

Be proficient in Python (for completing project)

Know basic NLP concept – To the extent that you understand concepts like train/dev/test set, model fitting, feature, supervised learning, etc. (We will not cover these in this course!)
Major Course Work

30% Homework Assignments
  15% One assignment
  15% One scribe

65% Final Project
  5% In-Class project presentation
  10% Literature review
  20% Experiment protocol
  5% Final poster presentation
  25% Project report

5% Participation (discussion in class or via Ed Discussion)
Scribe

Please sign up for one slot for this course (one week to complete your scribe).

Scribe template will be released by TA soon.

3 pages (excluding references) for a one-person team
5 pages if two people team together

Expectation: a coherent summary of lecture content, with in-depth coverage of technical details and discussions.
Assignment

Will be released on Apr 12th, due on May 1st
2~3 problem sets related to course content
Course Project

10% Literature review
20% Experiment protocol
5% Final poster presentation
25% Project report

Project Scope (there will be a separate lecture on this):  

One key element to justify: what is the human-centered aspect in your project?
Case studies of human factors in existing NLP/AI systems
New mythologies tailored to a human-centered problem
Position papers or a critic (talk to us first)
Applying a computational text-based method to real-world problem for social good
In-Person Expectation

This is an in-person class by default, but if you cannot come, **there’s a zoom option on Canvas**. This link will not be on the webpage to avoid zoom boom :}

Computing Credits

We provide two types of credits (stay tuned for emails on how to access):

• Computing credit on Google Cloud

• Credits for accessing foundation models
Late Days

• Each student will have a total of 4 free late (calendar) days. Final project papers cannot be turned in late under any circumstances.

• Once these late days are exhausted, any work turned in late will be penalized 10% per late day.

• If a group's assignment is late n days, then each group member is charged n late days.

• Late days are never transferrable between students, even students in the same group.
Breadth Requirement

Area C: Applications in AI, HCI
People and Society

For others, check with us!