CS329X: Human Centered NLP

Culture and Values

Diyi Yang
Stanford CS
Lecture Overview

✦ Cross-culture NLP
  Slides credit to Daniel Hershcovich, EACL workshop on CrossCultural Considerations in NLP
✦ Value Alignment in LLMs
✦ Values in ML Research

Slides credits to:
• Camille Harris on Values in ML Research
• Daniel Hershcovich on Cross-culture NLP
What is culture?

shared beliefs, values, practices, behaviors, and artifacts

a system of symbols

characterizes a social group

learned, not innate

integrated with daily experience

taken-for-granted notions, rules, moralities, and behaviors

dynamic process through which social meanings are constructed

Cultural anthropology: Asking questions about humanity (Welsch and Vivanco, 2015)
Take a closer look at culture

**Linguistic variation**
- Idiolect
- Sociolect
- Dialect
- Standardized language
- Language groups and families

**Cultural variation**
- Individual, personality
- Social group
- Geographical region
- Country, national culture
- Global cultural community
Cultural dimensions in NLP

Challenges and Strategies in Cross-Cultural NLP (Hershcovich et al., ACL 2022)
Give me a one-sentence **kale chips** recipe in English, Hebrew, Arabic, Danish, Mandarin Chinese, Korean, Greek, Hindi, Bulgarian and Turkish. Do not say which languages they are.

Linguistic form

Kale chips recipe: Preheat the oven to 350°F, tear kale into bite-sized pieces, toss with olive oil and salt, and bake for 10-15 minutes until crispy.

Kale chips recipe in Hebrew:

מתוўר ל-180 מעלות, קורא את עצו, חיפש זעמה או סלט, ואתה בHomeAsUp ל-10-15 דקות עד שיבוע, אוף בזינה:

Kale chips recipe in Arabic:

الخضروات الورقية: اشترى الخضروات الورقية، قم بقطعها إلى قطع صغيرة، ثم اضف زيت الزيتون والملح، واخبزها لمدة 10-15 دقيقة حتى تصبح مقليمة.

Kale chips recipe in Danish:

Kale chips opskrift:

Forvarm ovnen til 180 grader, rød kale i små bidder, bland med olivenolie og salt, og bag i ovnen i ca. 15 minutter.

Kale chips recipe in Chinese:

羽衣甘蓝脆片食谱：将烤箱预热至摄氏180度，将羽衣甘蓝撕成小块，加入橄榄油和盐调味，烤10-15分钟，直到变脆。

Kale chips recipe in Korean:

가레이 샐러드 레시피: 오븐을 180도로 예열하고 가레이를 작은 조각으로 쪼개 웨이터 오일과 소금과 섞어 15분간 구워서 바삭하게 만든다.

Kale chips recipe in Greek:

Συνταγή για τα κύπελλα: Τροποποιήστε το φούρνο στα 180 βαθμούς, κόψτε τα φύλλα κέλλι σε μικρά κομμάτια, ανακατεύτε με ελαιόλαδο και αλάτι, και ψήνετε για 10-15 λεπτά μέχρι να γίνουν τραγανά.

Kale chips recipe in Hindi:

 Scouts रेसिपी: ऑवन को 350 फाेरीहाइट पर पूर्व गाम करें, केल को छोटे टुकड़ों में दुकाएं, ऑलिव ओयल और नमक के स
Style

**BRAZILIAN-PORTUGUESE**

*Informal*

saiam disso, força de vontade!!
get out of it, willpower!!

*Formal*

Abandonem essa situação, tenham força de vontade.
Abandon this situation, have willpower!

**FRENCH**

*Informal*

Il avait les yeux braqués ailleurs.
He had his eyes fixed elsewhere.

*Formal*

Il ne prêtait pas attention à la situation.
He was not paying attention to the situation.

**ITALIAN**

*Informal*

in bocca al lupo!
good luck!

*Formal*

Ti rivolgo un sincero augurio!

---

Olá, Bonjour, Salvel! XFORMAL: A Benchmark for Multilingual Formality Style Transfer (Briakou et al., NAACL 2021)

---

Few-shot Controllable Style Transfer for Low-Resource Multilingual Settings (Krishna et al., ACL 2022)
Common ground

Adapt the recipe to the **common ingredients, cooking methods and tools** in each culture's cuisine.
Conceptualization

*Visually Grounded Reasoning across Languages and Cultures* (Liu et al., EMNLP 2021)
Snack

Grocery Store
Commonsense

Geo-Diverse Concepts

- Color of wedding dress
- Shower time
- Staple food
- Driver seat side
- Unit of measurement
- …

*GeoMLAMA: Geo-Diverse Commonsense Probing on Multilingual Pre-Trained Language Models* (Yin et al., EMNLP 2022)

*Figure 1: Average day time for the greeting good morning in different countries (USA, Brazil, Spain and India)*

**Grounding the Semantics of Part-of-Day Nouns Worldwide using Twitter** (Vilares & Gómez-Rodríguez, PEOPLES 2018)

*Good Night at 4 pm? Time Expressions in Different Cultures* (Shwartz, Findings 2022)
Give me one-sentence recipes for different **culture-specific dishes** in the following languages (without specifying the dish names): English, Hebrew, Arabic, Danish, Mandarin Chinese, Korean, Greek, Hindi, Bulgarian and Turkish

English: Season and grill marinated beef skewers to make a traditional BBQ dish.
Hebrew: Mix chickpeas, tahini, lemon juice, and garlic to make a creamy hummus dip.
Arabic: Stuff grape leaves with a mixture of rice, ground meat, and spices for a delicious and traditional appetizer.
Danish: Toast slices of rugbrød (rye bread) and top with sliced meats and cheese for an open-faced sandwich.
Mandarin Chinese: Fry diced chicken with ginger and garlic, then stir-fry with vegetables and soy sauce for a classic Chinese dish.
Korean: Cook rice cakes in a spicy sauce with vegetables and meat to make a traditional tteokbokki dish.
Greek: Season and grill lamb chops, then serve with a side of lemon potatoes and tzatziki sauce.
Hindi: Mix mashed potatoes with spices and shape into small balls, then fry to make spicy aloo tikki.
Bulgarian: Mix minced pork with onions and spices, then shape into sausages and grill for a traditional Bulgarian dish.
Turkish: Fry ground beef with onions and spices, then stuff into phyllo dough to make börek, a savory pastry.
Topics and entities

Contrasts. EN: rein [in debt], saddle [with debt], cap [debt], breed [money], gut [budget], [debt] hit, tackle [debt], boost, slow, drive, sluggish [economy], spur
Contrasts. ES: sangrar [dinero], ampliar, disminuir [la economía], superar [la tasa], emitir [deuda]

Figure 6: Contrasts identified by the model in NEWS.

Detecting Cross-Cultural Differences Using a Multilingual Topic Model (Gutiérrez et al., TACL 2016)

Dataset Geography: Mapping Language Data to Language Users (Faisal et al., ACL 2022)
Objectives and values

For each of these cultures, give a specific reason for why they would even want to make kale chips

- English: Kale is a common ingredient in British cuisine, and making kale chips could be a way to enjoy a crunchy and healthy snack.
- Hebrew: Many Israelis enjoy incorporating healthy and nutritious foods into their diets, and kale chips could be a way to do so while also adding some variety to their snacking options.
- Arabic: In Middle Eastern cuisine, kale is often used in stews or as a garnish, but making kale chips could be a new and interesting way to prepare the vegetable.
- Danish: Danish cuisine often features hearty, warming dishes, and kale chips could provide a lighter and more refreshing option for snacking or as a side dish.
- Mandarin Chinese: In China, many people prioritize health and wellness, and kale chips could be seen as a tasty and convenient way to get some extra nutrients into their diet.
- Korean: Korean cuisine features a lot of pickled and fermented vegetables, and making kale chips could be a fun and flavorful way to experiment with a different kind of vegetable preparation.
- Greek: Greek cuisine is known for its use of fresh herbs and vegetables, and incorporating kale into traditional Greek recipes or making kale chips could be a way to add some variety to the vegetable dishes.
- Hindi: Many traditional Indian dishes feature leafy greens such as spinach, but making kale chips could provide a new way to enjoy a different kind of green vegetable as a snack or side dish.
- Bulgarian: Bulgarian cuisine often includes grilled or roasted vegetables, and making kale chips could be a tasty and easy way to add some variety to these types of dishes.
- Turkish: Turkish cuisine is known for its use of herbs and spices, and making kale chips with some of these traditional Turkish seasonings could be a flavorful and healthy snack option.

Reasons

- Variety
- Sustainability
- Enjoyment
- Health
Moral Values

Hofstede’s Cultural Dimensions Theory
- Power distance
- Individualism/collectivism
- Uncertainty avoidance
- Masculinity/femininity
- Long-term orientation/short-term orientation
- Indulgence/restraint

World Values Survey
- Traditional/secular-rational
- Survival/self-expression

Moral Foundations Theory
- Care/harm
- Fairness/cheating
- Loyalty/betrayal
- Authority/subversion
- Sanctity/degradation
- Liberty/oppression

European Values Study
- Personal freedom
- Individual autonomy
- Social solidarity
- Ethnic tolerance
- Civic honesty
- Gender equality
- Liberal democracy

Cultures and Organizations: Software of the Mind (Hofstede, 1991)
Value bias in language models

Values are altered to reflect US culture

"1. I am in favor of limiting immigration. 2. I am in favor of limiting immigration for humanitarian reasons. 3. I am in favor of limiting immigration for economic reasons."

GPT-3 summarize

The Ghost in the Machine has an American accent: value conflict in GPT-3 (Johnson et al., arXiv 2022)
Exercise: how would culture shape a NLP task?

- Email writing assistant
- Sentiment analysis
- Chatbot
Adaptation of models

- Culturally grounded objectives
- Transfer learning
- Cultural prompts
- Knowledge representation
- Inference of desired target culture
Culturally grounded objectives

Data
- Selection
- Annotation
- Projection

Models
- Training
- Transfer
- Pre-training

Tasks
- Reflecting on meta-objectives
- Co-design and stakeholder agency
- Cross-disciplinary collaboration
Transfer Learning

Cross-Cultural Similarity Features for Cross-Lingual Transfer Learning of Pragmatically Motivated Tasks (Sun et al., EACL 2021)

Cross-Cultural Transfer Learning for Chinese Offensive Language Detection (Zhou et al., C3NLP 2023)
Cultural prompts: interviewing bots

Large Pre-trained Language Models Contain Human-like Biases of What is Right and Wrong to Do (Schramowski et al., Nature Machine Intelligence 2022)

Speaking Multiple Languages Affects the Moral Bias of Language Models (Hämmerl et al., arXiv 2022)

Probing Pre-Trained Language Models for Cross-Cultural Differences in Values (Arora et al., C3NLP 2023)

Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study (Cao et al., C3NLP 2023)
Commonsense Probing

GeoMLAMA: Geo-Diverse Commonsense Probing on Multilingual Pre-Trained Language Models (Yin et al., EMNLP 2022)
Cultural Translation

Entities

„Ich habe Merkel gesehen, als sie im ICE einen Berliner von Dietsch gegessen hat“

I saw Biden eating a Boston Cream from Dunkin’ Donuts on the Acela

Recipes

凉拌秋葵

Ingredients
- 8 oz (225 g) okra
- 2 teaspoons light soy sauce (or soy sauce)
- 1/2 teaspoons green Sichuan pepper oil (or more to taste)

Instructions
- Bring a medium pot of water to a boil. Add 1 teaspoon vegetable oil and a pinch of salt...
Case Study: the use of value-targeted dataset
Adapting Language Models to Society (PALMS)

Main idea: adjust the behavior of a pertained language model to be sensitive to predefined norms with value-targeted datasets

Key steps:
- Choose sensitive topics
- Describe the language model’s desired behavior
- Write prompts with value-targeted question-answer pairs

Adapting Language Models to Society (PALMS)

Sensitive Topics

Abuse, violence
Human body image
Mental health
Relationships
Terrorism
Protected groups
...

Desired Behavior

*Human characteristics and behavior*

**Position:** Oppose unhealthy beauty or likeability standards; support goodness, attractiveness, and likeability in humans being subjective

**Reasoning:** Human characteristics, such as beauty, and interpretation of behavior, such as “normality”, are subjective and cultural. Promoting one standard of normality or goodness risks reinforcing unhealthy, unattainable, or imperialist ideals
Adapting Language Models to Society (PALMS)
Domains

Linguistic Form and Style

- Literature
- Textbooks
- Recipes
- City plans

Objectives and Values

- Opinion pieces
- Political speeches
- Reviews
- Travel guides

Common Ground

- DIY guides
- Navigation

Aboutness

Topics; what information people care to convey in language
Who decides what is the desired target culture?
Building Culturally Aware AI?

- LMs are not culturally universal, but they can be adapted
- We can use LMs to bridge between cultures
- We must think outside the NLP box to do these things
Lecture Overview

✦ Cross-culture NLP
✦ *Values* in ML Research

---

The Values Encoded in Machine Learning Research

_Abeba Birhane*
abeba@mozillafoundation.org
Mozilla Foundation & School of Computer Science, University College Dublin, Dublin, Ireland

_Pratyusha Kalluri*
pkalluri@stanford.edu
Computer Science Department, Stanford University, Palo Alto, USA

_Dallas Card*
dalc@umich.edu
School of Information, University of Michigan, Ann Arbor, USA

_William Agnew*
wagnew3@cs.washington.edu
Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, USA

_Ravit Dotan*
ravit.dotan@berkeley.edu
Center for Philosophy of Science, University of Pittsburgh, Pittsburgh, USA

_Michelle Bao*
baum@stanford.edu
Computer Science Department, Stanford University, Palo Alto, USA

---
Develop an understanding of the values in ML research

• ML research, and scientific research in general, is often assumed to be **value-neutral** but this claim is typically not well supported or investigated.

• Value as defined in philosophy of science research

• A "value" of an entity is a property that is considered desirable for that kind of entity, e.g. regarded as a desirable attribute for machine learning research.
The Values Encoded in Machine Learning Research

- Developed annotation scheme for identifying values in research papers
- Extract values from 100 highly cited ML papers from 2008-2009, and 2018-2019
- Manual sentence-by-sentence textual analysis to determine presence of and motivations behind different values
- Qualitative analysis of affiliations and funding sources for these papers
Annotation Procedure

• Identify justificatory chain for motivation of papers
• Identify values upheld within each sentence
• Categorize the papers’ discussions of societal impact
• Document and categorize the author affiliations and funding sources
• Societal impacts of research are rarely mentioned
• Papers that do connect their work to a societal need do so very loosely, often without clear justification
• Connection to societal needs is generally only mentioned in the beginning of papers

<table>
<thead>
<tr>
<th>Justificatory Chain</th>
<th>% of Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does not mention societal need</td>
<td>68%</td>
</tr>
<tr>
<td>States but does not justify how it connects to a societal need</td>
<td>17%</td>
</tr>
<tr>
<td>States and somewhat justifies how it connects to a societal need</td>
<td>11%</td>
</tr>
<tr>
<td>States and rigorously justifies how it connects to a societal need</td>
<td>4%</td>
</tr>
</tbody>
</table>
She’s Going Soft! – A commentary on “hard” and “soft” sciences

Which one is “harder”

A. Improving RLHF to make LLMs more aligned
B. Tracking misinformation online
C. Designing an interactive email writing assistant
The truth is that any scientific discipline When practiced properly When studied rigorously Is hard
Negative Potential

- Negative potential is rarely discussed, only 2 papers in the sample
- In those 2 papers, negative potential was mainly discussed hypothetically and generally, rather than discussing the specific negative potential of the presented model/concept
- Even in applications with well known negative potential such as surveillance and misinformation technologies do not discuss negative impacts

<table>
<thead>
<tr>
<th>Discussion of Negative Potential</th>
<th>% of Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does not mention negative potential</td>
<td>98%</td>
</tr>
<tr>
<td>Mentions but does not discuss negative potential</td>
<td>1%</td>
</tr>
<tr>
<td>Discusses negative potential</td>
<td>1%</td>
</tr>
<tr>
<td>Deepens our understanding of negative potential</td>
<td>0%</td>
</tr>
</tbody>
</table>
Top Values in ML Communities

• Performance (96% of papers),
• Generalization (89%),
• Building on past work (88%),
• Quantitative evidence (85%),
• Efficiency (84%)
• Novelty (77%)
Performance

• Typically measured as correctness averaged across predictions
  • Typically equal weighting for all predictions
  • Fairness research considers alternatives
• Pre-established large datasets with discrete ground truth labels are typically used for evaluation
  • Choice of dataset typically based on improving on baselines from prior work
  • Choice of dataset also justified by real world applicability
  • Using the same dataset encodes favoritism when the dataset isn’t accessible to all.
• Overall performance is highly valued
  • So much so that “success” and “improvement” become synonymous with accuracy
Generalization

- Shown by performing well on multiple tasks or multiple datasets
- The choice of datasets or choice of task indicating generalization are often not justified, unclear when performance will continue to generalize beyond carefully selected tasks or datasets
- Generalization generally is not with respect to more realistic applications
- Assuming generalizability often assumes new data should be treated similarly to previously seen data
Efficiency

- Efficiency commonly measured with respect to the ability to scale
- Not typically measured with respect to saving resources, in fact, papers that emphasize scalable ‘efficiency’ often require more resources than previous applications
- This incentivising scaling without consideration for resources incentives the most powerful actors to make increasingly inaccessible and reducing the ability to compete
Novelty & Building on Past Work

- Technical novelty is most valued
- As opposed to applying existing technology to a novel domain or novel philosophical argument
- Minimal critique of prior work
- Limitations of prior work are typically pointed out to emphasize the contributions of the current work
- Papers are rarely explicitly critical of prior work beyond pointing out these limitations
- Works that critique or rectify socially harmful aspects of prior work/datasets/goals are typically not seen as novel
• In 2008/09, 24% of the top cited papers had corporate affiliated authors & in 2018/19 this statistic almost tripled to 71%.

• A concentration of a few large tech firms increasing more than fivefold, from 11% to 58%.

Figure 2: Corporate and Big Tech author affiliations.
User Rights and Ethical Values

Values like privacy, user input or influence, avoiding social biases, justice, and fairness are reflected in less than 10% of the papers analyzed.

Negative potential of applications discussed in 1% (2 papers)
Positive Directions for User Rights and Ethical Values

Increase in value driven research organizations that center on ethical values

Increase in research funded by nonprofits

Increasing academic, nonprofit, and grassroots interest in upholding ethical values
Discussion Questions

1. Under what circumstances do the conventional values in the field contradict ethical values like privacy, justice, and combating social bias?

2. What are the positive and negative implications of the increasing influence of tech companies of funding and conducting academic research in the field?

3. What are the implications of ~50 universities producing most of the most citing work?
Fireside Chat with Pratyusha Kalluri