Challenges and Strategies in Cross-Cultural NLP Reading Presentation

Jinzuomu Zhong Osman Batur İnce



The University of Edinburgh CDT in Designing Responsible NLP Designing Responsible Natural Language Processing

§1 Introduction

A Framework for Culture Awareness

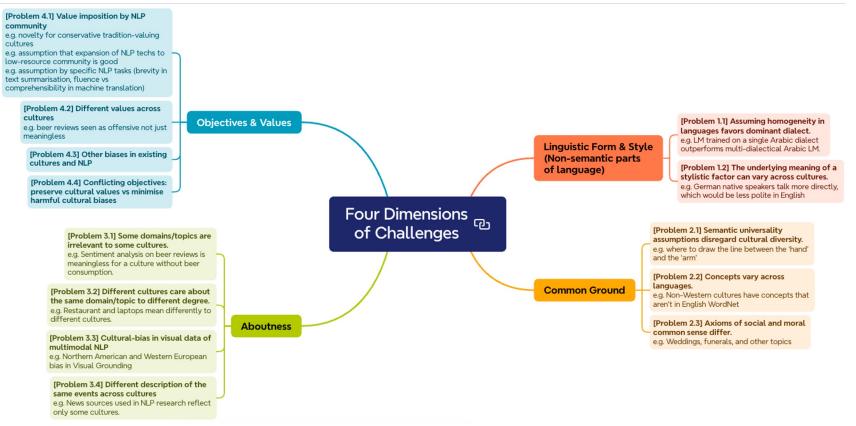
- Motivation
 - Recent progress on multi-linguality
 - Multilingual models (mBERT, mT5, Aya)
 - Linguistic diversity!
- What about culture?
 - Culture
 - material and non-material aspects
 - NLP systems should not be culturally maladapted!

§1 Introduction

A Framework for Culture Awareness



§2-5 Four Dimensions of Challenges



Element 1 - Linguistic Form & Style

Linguistic Form & Style (Non-semantic parts of language) [Problem 1.1] Assuming homogeneity in languages favors dominant dialect. e.g. LM trained on a single Arabic dialect outperforms multi-dialectical Arabic LM.

[Problem 1.2] The underlying meaning of a stylistic factor can vary across cultures. e.g. German native speakers talk more directly, which would be less polite in English

- "[T]reating "a language" as a homogeneous mass limits cultural adaptation, and runs the risk of privileging certain cultures over others."
 - any examples?

- Differences in form across different English speaking ethnicities in depression diagnosis.
 - Can language models capture these variations?

Element 2 - Common Ground

[Problem 2.1] Semantic universality assumptions disregard cultural diversity. e.g. where to draw the line between the 'hand' and the 'arm'

[Problem 2.2] Concepts vary across languages. e.g. Non-Western cultures have concepts that aren't in English WordNet

[Problem 2.3] Axioms of social and moral common sense differ. e.g. Weddings, funerals, and other topics

- Can one NLP model capture the common ground of different cultures, or do we need separate models for each culture?
- Hand vs arm: <u>https://wals.info/feature/129A#2/0.0/178.8</u>
- Finger vs hand: <u>https://wals.info/feature/130A#2/14.9/153.8</u>
- Number of colo(u)rs: <u>https://wals.info/feature/133A#2/32.5/151.7</u>

Common Ground

Element 3 - Aboutness

- Simple message: Different cultures care about different things, in different ways.
- Aboutness vs Common Ground
 o difference?
 - overlap? causal relation?
- Harms of NLP not caring about aboutness?

[Problem 3.1] Some domains/topics are irrelevant to some cultures.

e.g. Sentiment analysis on beer reviews is meaningless for a culture without beer consumption.

[Problem 3.2] Different cultures care about the same domain/topic to different degree. e.g. Restaurant and laptops mean differently to different cultures.

[Problem 3.3] Cultural-bias in visual data of multimdal NLP e.g. Northern American and Western European bias in Visual Grounding

[Problem 3.4] Different description of the same events across cultures e.g. News sources used in NLP research reflect only some cultures.

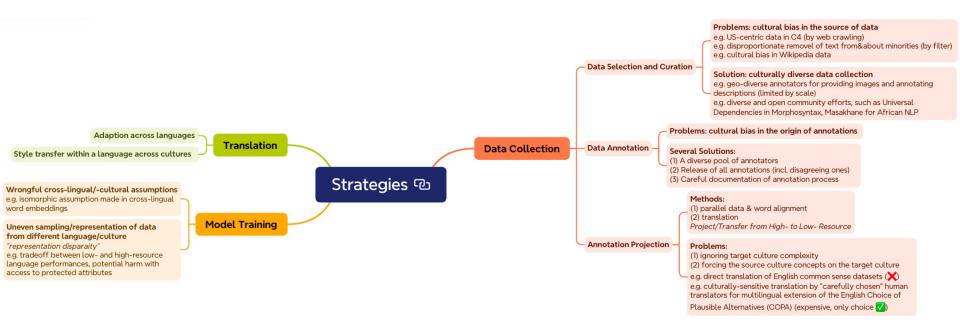
Aboutness

Element 4 - Objectives & Values

- Simple message: Different cultures have different values.
- Other biases vs Culture biases
 - difference? relations?
 - "[C]ountering bias in NLP can be seen as shifting existing cultures into hopefully more equitable ones."
- Preserve vs Intervene?
 - other examples?

5	[Problem 4.1] Value imposition by NLP community e.g. novelty for conservative tradition-valuing cultures e.g. assumption that expansion of NLP techs to low-resource community is good e.g. assumption by specific NLP tasks (brevity in text summarisation, fluence vs comprehensibility in machine translation)		
		[Problem 4.2] Different values across cultures e.g. beer reviews seen as offensive not just meaningless	Objectives & Values
		[Problem 4.3] Other biases in existing cultures and NLP	
		[Problem 4.4] Conflicting objectives: preserve cultural values vs minimise harmful cultural biases	J

§6 Three Strategies



Strategy 1 - Data Collection

	Data Selection and Curation Problems: cultural bias in the source of data e.g. US-centric data in C4 (by web crawling) e.g. US-centric data in C4 (by web crawling) e.g. disproportionate removel of text from&about minorities (by fire.g. cultural bias in Wikipedia data Solution: culturally diverse data collection e.g. geo-diverse annotators for providing images and annotating descriptions (limited by scale) e.g. diverse and open community efforts, such as Universal Dependencies in Morphosyntax, Masakhane for African NLP	lter)
Data Collection	Data Apposition	atasets (X) hosen" human sh Choice of 10
	Annotation Projection – (1) ignoring target culture complexity	
	 (2) forcing the source culture concepts on the target culture e.g. direct translation of English common sense datasets (×) e.g. culturally-sensitive translation by "carefully chosen" human translators for multilingual extension of the English Choice of Plausible Alternatives (COPA) (expensive, only choice ✓) 	

Strategy 2 & 3 - Training & Translation

- Similar Problems
 - Translation vs Projection?
 - Representation (same as in Data)
- Algorithmic/Technical Approach
 - Modelling assumptions
 - Sampling techniques
 - Careful translation as assistance



Translation

Style transfer within a language across cultures

Wrongful cross-lingual/-cultural assumptions

e.g. isomorphic assumption made in cross-lingual word embeddings

Uneven sampling/representation of data from different language/culture

"representation disparaity" e.g. tradeoff between low- and high-resource language performances, potential harm with access to protected attributes

Model Training

§7 Conclusion

• Impact of Culture on NLP

Cultural knowledge and values affect NLP

Beyond Western Cultures

- Represent non-Western societies
- Let non-Western societies express themselves

Decolonising NLP

• Move away from homogenous approaches

On Writing

- Title: Challenges and Strategies in Cross-Cultural NLP
 - Comments?
 - What is Cross-Cultural NLP?
- Chapter/Section Titles
 - Good! Avoid meaningless "Introduction/Method/Results/..."
- Macro-/Micro- structure
 - Comments?

Let's Be Critical

- 1. Lack concrete NLP examples for stylistic variations
- 2. Lack of benchmarks for evaluating NLP systems
- 3. No clear examples of NLP models making cultural mistakes
- 4. Cross-cultural vs cross-lingual?

Further Questions/Comments?