Title: Building Knowledge-Guided Lexica to Model Cultural Variation
Authors: Shreya Havaldar , Salvatore Giorgi , Sunny Rai , Young-Min Cho , Thomas Talhelm, Sharath Chandra Guntuku , & Lyle Ungar
Published in: NAACL 2024
Link: https://aclanthology.org/2024.naacl-long.12/

Topic, Research Question, Contribution

In this paper, the authors investigate how to create a knowledge-guided lexica. A lexica is a curated set of words. To demonstrate how their knowledge-guided lexica works, they chose to study individualism and collectivism in the United States. The authors present the following research problem for the NLP community, and then use their knowledge-guided lexica to start answering the question: "How do we measure variation in cultural constructs across regions using language?" Their main contribution is: "We develop knowledge-guided lexical models and demonstrate their ability to measure individualism and collectivism. Our method (1) is highly scalable, (2) encodes domain knowledge from cultural psychology, and (3) does not require additional labeled data."

Challenges

There are two main challenges to building a *good* knowledge-guided lexica. First, building lexicons based on cultural theory normally takes a long time. Second, many lexicons contain words that correlate negatively with the construct they are trying to measure.

When evaluating their lexicon, one challenge was county interpolation. Twitter data is not available for every county, so they use a Gaussian Process regression model to interpolate across counties.

Methodology

Step 1: they ask an expert to develop a small set of seed words. Step 2: they use word embeddings to expand the number of words, looking for synonyms and words similar to the concept. Step 3: they filter out words that are rare or correlate negatively with other words in the lexicon.

Evaluation, Evidence, Statistics

To evaluate how well their lexicon works, they calculate the weighted frequencies of each lexicon word in the Twitter corpus and then compare the results with previous research on collectivism and individualism to see if their metric is correlated with the existing research. They use pairwise product-moment correlations.

Results

When the researchers evaluated their lexicon, they found that the expansion step is very important. However, the thresholds for expansion are different for different concepts. Contrastingly, the threshold for purification does not seem to matter as long as it is positive. They also find that their model correlates with the previous research on collectivism and individualism better than seed words only or zero-shot GPT-3.5. _