Annotating and Training for Population Subjective Views

Maria Alexeeva[1] Caroline Campbell Hyland[1] Keith Alcock[1] Allegra A. Beal Cohen[2]
Hubert Kanyamahanga[3] Isaac Kobby Anni[3] Mihai Surdeanu[1]

[1]University of Arizona, [2]University of Florida,
[3]International Crops Research Institute for the Semi-Arid Tropics

Overview

- We release a dataset for identifying subjective views of individuals or groups that can potentially impact their behavior, e.g.:

Consumers generally ***recognize that cheaper prices correspond with lower quality*** *and tend to remain loyal to their preferences when prices increase.*

cheaper prices = lower quality
↓
remaining loyal to preference with prices increasing

- We train a model for identifying such subjective views in text using the created dataset.
- We discuss ways in which we mitigate issues related to human label variation (Plank, 2022) and provide support for embracing it.

Defining the Annotation Task

- Subjective views = beliefs + attitudes:
  - **Belief:** *Consumers generally recognize that cheaper prices correspond with lower quality*
  - **Attitude:** *However, the members cultivated rice twice in 2009/10 [...] because they did not plan to cultivate rice in 2010/11 and wanted to secure a whole year's worth of rice for their own consumption.*
- Reported vs. author beliefs
  - **Reported:** *In Germany, they think that they may have risked to much [...]*
  - **Author:** *We here in Germany think that we may have risked too much [...]*
- Actually held vs. hypothetical beliefs
  - **Actually held:** *Rice production is considered a supplementary, non-commercial activity in the region.*
  - **Hypothetical:** *If local actors perceive too much initial risk to invest in their own brands [...]*
- Complete vs. incomplete beliefs
  - **Complete:** *Red-billed Quelea is considered the most numerous bird world- wide with population numbers totaling about 1.500 million [...]*
  - **Incomplete:** *It is considered the most numerous bird world- wide with population numbers totaling about 1.500 million [...]*
- Compare to stance detection (Mohammad et al. 2016), opinion mining (Wankhade, Rao, and Kulkarni 2022), and beliefs in Tracey et al. 2022

Dataset Statistics

- Train and test partitions
  - **Train:** Mechanical Turk + the team quality control
  - **Test:** annotated by the domain expert and the team
- Known vs. unknown triggers:
  - **Known:** a set created in collaboration with domain experts and supplemented by the team based on data analysis
  - **Unknown:** belief triggers potentially present in the dataset, but not identified in advance

Measure	Train			Test	
	known triggers	unk. triggers	unk. in training	known triggers	unk. triggers
N documents	59	65	65	50	43
N data points	1044	9769	1440	400	193
N positive class	360	0*	0*	202	12
% positive class	34%	0*	0*	50.5%	6%
Unique triggers	95	N/A	N/A	72	12

Annotation Issues

- Sentence ambiguity: *[...] and some farmers apply urea (called 'salt'), saying that leaf color becomes healthy.*
- Context ambiguity: *I therefore consider that the global rice VC is part of the context, and I do not make it the focus of the research.* ***Nevertheless, [importers] are considered as part of the domestic VC.***
- Human factor: focus on different aspects of the guidelines, bad faith annotation.

Experiments

The task: provide a binary label indicating whether or not a given sentence contains a belief.

Evaluation: in-domain (training cross-validation) and out-of-domain (test partition)

- The main model:

Model	P	R	F1
In-domain	$0.68_{\pm0.05}$	$0.73_{\pm0.07}$	$0.7_{\pm0.02}$
Out-of-domain	$0.77_{\pm0.03}$	$0.80_{\pm0.03}$	$0.78_{\pm0.02}$

- MTurk Annotation Threshold:

Partition	Setting	P	R	F1
CV	MTurk0.5	$0.72_{\pm0.06}$	$0.82_{\pm0.08}$	$0.76_{\pm0.02}$
	MTurk1.0	$0.41_{\pm0.07}$	$0.49_{\pm0.09}$	$0.44_{\pm0.04}$
	MTurkQC	$0.68_{\pm0.05}$	$0.73_{\pm0.07}$	$0.7_{\pm0.02}$
Test	MTurk0.5	$0.54_{\pm0.03}$	$0.87_{\pm0.02}$	$^{*}0.67_{\pm0.02}$
	MTurk1.0	$0.54_{\pm0.04}$	$0.42_{\pm0.03}$	$0.47_{\pm0.03}$
	MTurkQC	$0.77_{\pm0.03}$	$0.8_{\pm0.03}$	$^{*}0.78_{\pm0.02}$

- Marked trigger: in-domain

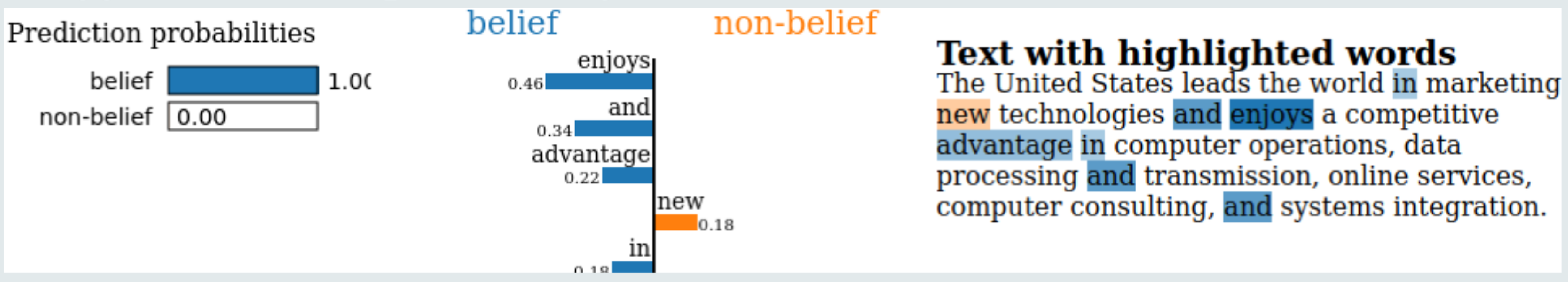
Model	P	R	F1
Unmarked trigger	$0.68_{\pm0.05}$	$0.73_{\pm0.07}$	$0.7_{\pm0.02}$
Marked trigger	$0.72_{\pm0.06}$	$0.72_{\pm0.05}$	$0.72_{\pm0.05}$

- Marked-trigger: out-of-domain

Model	P	R	F1
Unmarked trigger	$0.77_{\pm0.03}$	$0.8_{\pm0.03}$	$0.78_{\pm0.02}$
Marked trigger	$0.81_{\pm0.03}$	$0.74_{\pm0.03}$	$0.77_{\pm0.02}$

Error Analysis

- False positives:
  - going against task assumptions (e.g., assigning positive class for incomplete beliefs)
  - triggers with multiple meanings:



- False negatives:
  - anti-modal-verb bias and anti-long-sentence bias:

  - anti-first-person bias:

- General observations:
  - The model learns to pay attention to triggers (e.g., *considered, believe*).
  - The model learns new triggers (e.g., *enjoy, problematic, likely*).

Acknowledgments

The authors thank the anonymous reviewers for helpful discussion. This work was supported by the Defense Advanced Research Projects Agency (DARPA) under the Habitus program. Maria Alexeeva and Mihai Surdeanu declare a financial interest in lum.ai. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies. The annotation work was partially supported through Research and Project (ReaP) Grant from the University of Arizona Graduate and Professional Student Council (GPSC).