Crowdsourcing Designs for Data Collection: An Experimental Analysis

Abstract

Data about political conflict are mostly coded from news reports and other secondary sources. Collecting such data is costly, and as a result many datasets are limited temporally or spatially, and updated infrequently. Existing research suggests that crowdsourcing offers an efficient alternative to more expensive, traditional data collection methods. However, researchers who wish to leverage the efficiencies of crowdsourcing will find themselves making many design decisions with little to no empirical guidance. Using the Militarized Interstate Dispute data, we experiment with crowdsourcing designs in an effort to improve the process by which quality conflict data is collected. We find that workers perform worse when provided a question tree structure, and that performance is greatly affected by the type of action being classified. Preliminary evidence suggests that automated answer cues increase performance, but only when the correct answer is in the list of suggested answers. Viewing the prior worker's response decreases performance.

Research Question

How can researchers optimally design crowdsourcing systems to produce the highest quality coding of event data?

Our Goal

Improve scalability of human classification for event data.

Application

Experiment with different survey structures to classify Militarized Interstate Dispute (MID) data from news stories.

Features of a MID

Actor 1 ➔ MID Action Type ➔ Actor 2

Research Design

- Sample of 290 documents, mostly from 2011
- Minimum 15 per MID action type
- Many contain multiple events
- Workers recruited from Mechanical Turk:
  - Answered questions pertaining to the highest level action in the story
  - Assigned to one of six experimental conditions
  - Paid 50 cents, plus 25 cent bonus if random question answered correctly
- Surveys hosted on Qualtrics

Experiment Conditions

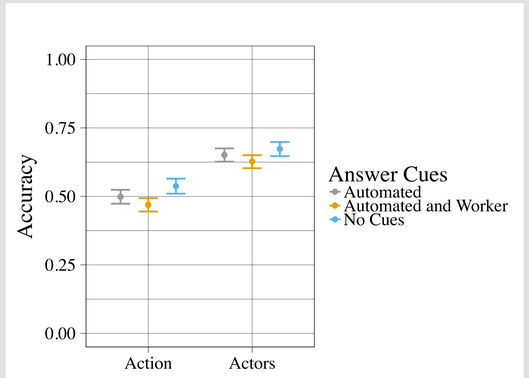
	IT (1)	INT (2)	ITNS (3)	INTNS (4)	WT (5)	WNT (6)
No additional information			✔	✔		
Automated suggestions	✔	✔			✔	✔
Prior worker's responses					✔	✔
Question tree structure	✔		✔		✔	
Flat question structure		✔		✔		✔

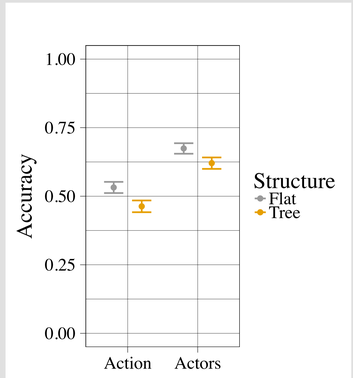
Experiment Details

- "Tree" breaks down the questions pertaining to the action type into simpler, mostly yes/no questions.
- "Flat" asks the worker to code the action type, and provides a list of all possible action types.

- Action type and specific locations are suggested using the Never Ending Language Learning pipeline.
- Initiator, target, and country location are suggested using Phoenix actor dictionaries.
- Dates are suggested using regular expressions.

Results for Action and Actors



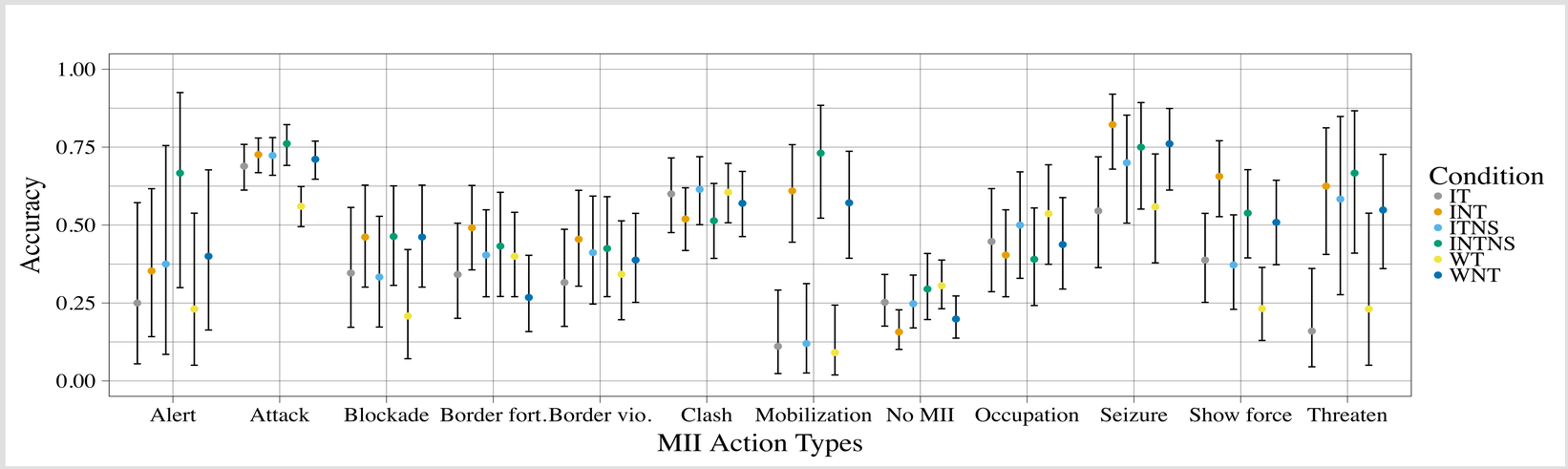


Conclusions & Future Results

- Providing more structure (via question tree) decreases accuracy
- Accuracy varies considerably by action type
- Automated answer cues increase performance, but only when the automated suggestion list contains the true answer

- How does test score impact accuracy? How can we identify and upweight "expert" workers?
- Other event components---Date, Location, etc.
- Does accuracy change when the correct answer is in the list of automated suggestions?
- Are workers able to identify documents with multiple incidents?

Results by Action Type



Vito D'Orazio*, Michael Kenwick, Matthew Kelly, Dennis Okyere, Glenn Palmer, David Reitter, Zhanna Terechshenko

dorazio@utdallas.edu

Supported by NSF SBE-SES #1528624 and #1528409