Sprout: Crowd-Powered Task Design for Crowdsourcing

Jonathan Bragg
University of Washington
(Stanford University)

Mausam
Indian Institute of Technology
Delhi

Daniel S. Weld
University of Washington
Crowdsourcing Task Platforms

Volunteer
- Zooniverse
- Galaxy Zoo
- FoldIt

Paid
- Amazon Mechanical Turk
- Upwork (formerly oDesk)
Instructions:

Is this a car?
Instructions:

Is this a car?
Instructions:

Is this a real (not cartoon) car?
Instructions:

Is this a real (not cartoon) car that a person can drive on a street?

Low Inter-annotator Agreement

Yes

No
Instructions:

Is this a single real (not cartoon) car that a person can drive on a street?
Instructions:

Is this a single real (not cartoon) car that a person can drive on a street? The whole car must be visible for it to count.
Iterative Task Design Loop

Perform (explore) the task
- Task definition changes
- Concept evolution
  [Kulesza et al., CHI ‘14]

Revise Instructions

Run (debug) the task
- Worker mistakes initially unknown
- Model mismatch / debugging

Revise Training & Testing

Low Inter-Annnotator Agreement
Task Design is a Major Problem

• Difficulty for Requesters; Importance for Work Quality
  • [Alonso & Mizzaro, Information Processing and Management ‘12; Papoutsaki et al. HCOMP ‘15; Liu et al., NAACL ‘16]

• Risk to Workers: Task Design Flaws & Unfair Rejections
  • [McInnis et al., CHI ‘16; Gadiraju et al., HT ’17; Wu & Quinn, HCOMP ‘17]
...and it matters

[Freelancing in America: 2017. Upwork and Freelancers Union]
How can we design tools for task design?

Perform (explore) the task
- Task definition changes
- Concept evolution
  [Kulesza et al., CHI ’14]

Run (debug) the task
- Worker mistakes initially unknown
- Model mismatch / debugging

Revise Instructions

Revise Training & Testing

Low Inter-Annotator Agreement
Confusions
- Multiple Cars
- Partial Car
- Railroad Car

Instructions
Is this a single real (not cartoon) car that a person can drive on a street?
The whole car must be visible for it to count.
**Sprout: Debugging-Prioritized Exploration**

**Confusions**
- Multiple Cars
  - [✓](#) [—even though it's marked, it's not a real car](#)
  - [▼](#) [because it's clearly not a real car](#)
- Partial Car
  - [✓](#) [even though it's not fully visible](#)
- Railroad Car
  - [✓](#) [even though it's not fully visible](#)

**Instructions**
Is this a single real (not cartoon) car that a person can drive on a street? The whole car must be visible for it to count.

---

**Crowd-Powered Debugging**

All No  ↔  All Yes
Confusions

- Multiple Cars
  - Partial Car
  - Railroad Car

Instructions

Is this a single real (not cartoon) car that a person can drive on a street? The whole car must be visible for it to count.

Test Q’s

Yes

No
Is this a single real (not cartoon) car that a person can drive on a street? The whole car must be visible for it to count.
Sprout Compiles Test Q’s into Training & Testing

Interactive Tutorial → Test Questions → Pass? → Yes → Task Questions

No

Gated Instruction [Liu et al., NAACL ‘16]
Your instructions for workers go here. Use twitter mention notation to reference items, for example, @444 refers to item 10 and will preview as. You may also use other types of Markdown to format your instructions, like

- This will be a list item
- This will be **bold**

**Write**

Is this an image of a car?
Select "yes" even if there are multiple cars (e.g., @444)

**Recommended test questions**

because you mentioned 444 in the instructions

**Test questions**

Yes

No
Requester User Study

• Goal: improve underspecified instructions prompts
• 2 domains from prior work (car images and travel websites)
• Baseline no-crowd interface: Structured Labeling [Kulesza et al. CHI 2014]
• Within-subjects (counterbalanced interface)
• 11 participants
  • varying crowdsourcing experience
  • graduate and undergraduate students
  • 5 male, 5 female, 1 other
Sprout preferred by requesters

“Which interface did you prefer for creating instructions?”

“[c]ategorizing the inputs, showing me the cases where there was confusion, etc., made it SUPER easy to identify cases that needed clarification.” (P1)

“The similarity metrics seem to be working great and the suggested items are great for testing the points I emphasized in the instructions.” (P2)

20-29% of test questions from suggestions
Sprout aids instruction comprehensiveness

• Longer instructions and more examples (mean=4.6 vs 2.6 on travel task)
• More resolved ambiguous categories (mean=4.0 vs. 2.8 on cars task)
  • 2 experts coded number of distinct categories
Some requesters wanted to use both tools

• Structured labeling benefits
  • No bias (P2)
  • “Let me try doing the task myself” (P3)

• Tradeoff: “It sucks that you have to start from a completely blank slate. [SPROUT] gave you some more support.” (P4)
Task Design Tool Landscape

Crowd-Powered Exploration Support

Revolt
[Chang et al., CHI ‘17]

Sprout

Limited Task Improvement Support

Structured Labeling
[Kulesza et al., CHI ‘14]

Daemo [Gaikwad et al., CSCW ‘17]
WingIt [Manam & Quinn, HCOMP ‘18]

Limited Exploration Support
Future Work

• Expanded roles for workers & algorithms
  • More automated instructions: Infer requester decisions (“legal precedent”)
  • Engaging, progressive tutorial
  • Interface design
  • Workflow design (task decomposition)

• More studies
  • Field deployments

• Beyond labeling tasks
  • Open-ended & creative tasks
Thanks!
https://jonathanbragg.com

Sprout: Crowd-Powered Task Design for Crowdsourcing

Jonathan Bragg
UW
(Stanford)

Mausam
IIT Delhi

Dan Weld
UW

NSF
ONR
WRF
Google
IBM
Bloomberg
Government of India