A Needle in a Haystack: An Analysis of High-Agreement Workers on MTurk for Summarization

Lining Zhang[1], Simon Mille[2], Yufang Hou[3], Daniel Deutsch[4], Elizabeth Clark[5], Yixin Liu[6], Saad Mahamood[7], Sebastian Gehrmann[5], Miruna Clinciu[8], Khyathi Chandu[9], João Sedoc[1]
[1]New York University, [2]ADAPT Centre, DCU, [3]IBM Research, [4]Google, [5]Google Research, [6]Yale University, [7]trivago N.V., [8]University of Edinburgh, [9]Allen Institute for AI

Overview

Motivation: design **two-step recruitment pipeline** of high-quality *Amazon Mechanical Turk (MTurk)* workers for **text summarization** through *Human Intelligence Task (HITs)* given:

- Automatic metrics: **problematic** sometimes
- Best practices for recruitment on MTurk for human evaluations: **poorly understood**

Contribution:

- establish a recruitment pipeline to **build a pool** of annotators with **high agreement**
- successfully recruit **12 out of 200 (6%)** superior annotators with **lower costs** for **large scale** tasks
- **match or surpass** the *inter-annotator agreement (IAA)* of **experts** and **statistical techniques** (further calibration required for **correctness**)

Pipeline Design

The pipeline comprises a **qualification task** and an **endurance task**, followed by a **reference-based task**.

Qualification Task
- 3 documents, 1 summary, 6 dimensions
- evaluate **multiple dimensions correctly**

Endurance Task
- 10 HITs, 1 document and 4 summaries, saliency
- capacity for handling **heavy workload**

Reference -based Task
- 30 HITs, 1 reference and 4 candidate summaries, information coverage
- tests the **general performance**

** For the qualification task, we conduct 4 rounds each with 50 MTurk workers with **statistical test** for stability.*

Experiments and Results

Stage 0: MTurk Qualification Settings

Pre-task qualifications on workers can be set:
- Location
- Number of HITs Approved
- HIT Approval Rate (%) for all Requesters' HITs, etc

Stage 1: Qualification Task

Components of Qualification Task.

Workers Categorization (4 types):
- GOLD: all correct + attention check passed
- SILVER: all but 1 correct + attention check passed
- BROZE: attention check passed
- BLOCK: attention check not passed

26 (8 GOLD, 18 SILVER) qualified workers **(13% of 200 participants)**.

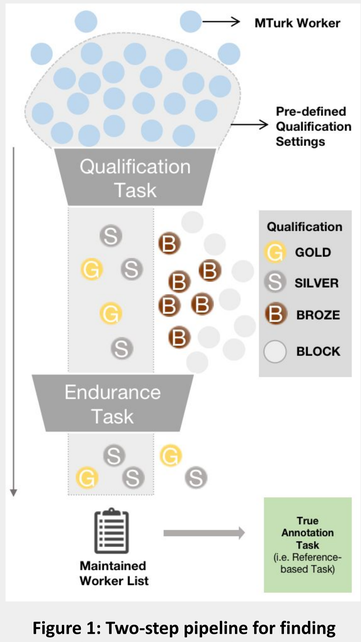

Figure 1: Two-step pipeline for finding high-agreement MTurk workers.

Stage 2: Endurance Task

- **12 (4 GOLD, 8 SILVER)** MTurk workers passed **(6% of 200 participants)**
- Best Cohen's Kappa: **0.55** (Across Groups)
- Best Krippendorff's Alpha: **0.443** (GOLD)

❖ **Comparison with experts:** ***Higher*** IAA
❖ **Detection of abnormal worker:** assign scores ***before*** the time for reading the document

Stage 3: Reference-based Task

Qualified Pipeline Workers:
- **8 (out of 12)** MTurk workers finished all HITs
- Best Cohen's Kappa: **0.68** (GOLD)
- Krippendorff's Alpha: **0.534** (all scores)

Baseline MTurk Workers:
- Krippendorff's Alpha (statistical filter-**MACE**): **0.380**
- **Incomplete** HIT coverage & **fewer** workers per HIT

CloudResearch MTurk Workers:
- Krippendorff's Alpha (high-quality platform): **0.513**
- **lower** task **acceptance rate**

	Pipeline	MACE(0.5)	CloudResearch
Num. of initial workers	200	276	45
% of workers kept	4%	19.2%	17.8%
HIT coverage	30/30	30/30	30/30
Avg. num. workers per HIT	8	2.4	8
Krippendorff's Alpha	**0.534**	0.380	0.513
Cost per worker (for Avg. num. workers per HIT)	**$27**	$175	$31

Table 1: Comparison between approaches of crowd annotators for the reference-based task.

Discussion:
Pre-task filtering of our pipeline:
- **avoid the waste** of time and resources
- achieve **high agreement** at a lower cost and a full coverage of HITs
- **similar quality** (Spearman's correlation) to CloudResearch

Correctness Analysis

We perform **analysis of correctness across annotation sources** on **50 random annotation** questions from the reference-based task.
- Pipeline and CloudResearch workers had a **significant Spearman's correlation**
- Pipeline may **not guarantee** the training of the **correctness**
- GPT models correlated well with **expert judgments**

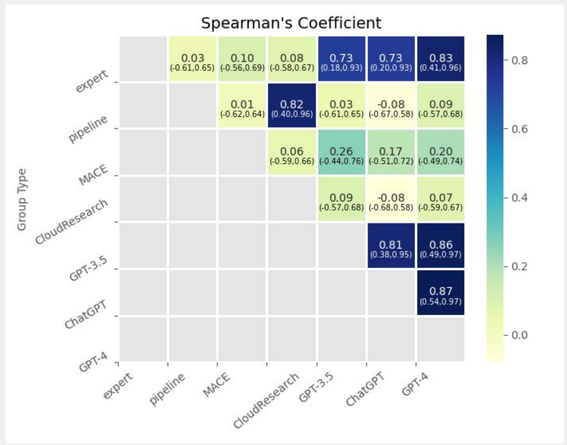

Figure 2: Spearman's coefficient with 95% confidence interval on 50 samples.

Conclusion and Limitations

Conclusion:
Serves as the **best practice** to:
- **high-agreement** annotations at large scale and lower cost
- **avoid resource waste** on discarded annotations

Limitations:
- **English summarization** on **MTurk** platform
- **designed questions** not "panacea" solutions
- **no guarantee** for the training of **correctness**

Reference:
The Perils of Using Mechanical Turk to Evaluate Open-Ended Text Generation (Karpinska et al., EMNLP 2021)
Learning Whom to Trust with MACE (Hovy et al., NAACL 2013)