Towards a Requester-centered Study on the Use of 'Bots' for Completing Tasks

Jonas Oppenlaender, Simo Hosio
University of Oulu
Oulu, Finland
{firstname.lastname}@oulu.fi

Abstract

Natural language models have become powerful tools for generating natural-sounding texts. The widespread interest in this technology has produced language models and tools that are available to the public. In the hands of "malicious workers," language models could be used to exploit tasks on crowdsourcing platforms for financial gains. This raises concerns about the potential impact of automated answer generation on the quality, reliability, and validity of data collected on crowdsourcing platforms, and on the future of crowdsourcing in general. In this work-in-progress, we present our efforts of studying related concerns among researchers in the crowdsourcing and human computation community.

Technological advances in the area of Machine Learning (ML) and Natural Language Processing (NLP) make it possible to generate natural-sounding text given little (few-shot) or no (zero-shot) examples as input. The availability and usability of natural language models has improved in recent years. Today, language models, such as , such as BERT, GPT-2, and GPT-Neo, are available as open source and can be executed with only a few lines of programming code.

Powerful natural language models will inevitably find their way into the toolkit of "malicious" crowd workers who exploit this technology for illicit uses. Malicious workers are "workers with ulterior motives, who either simply sabotage a task or try to quickly attain task completion for monetary gains" (Gadiraju et al. 2015). Malicious workers are likely to leave "untrustworthy" and "dishonest answers" that are either random, artificially generated, or copied from another source (Difallah, Demartini, and Cudré-Mauroux 2012; Gadiraju et al. 2015).

In the hands of malicious workers, the use of automated or semi-automated tools for answering tasks is a concern for requesters and a threat to the quality of data collected on crowdsourcing platforms as well as the research field of crowdsourcing and human computation as a whole. Crowdsourcing platforms are marketplaces for subjective human insights and judgments exercised by crowd workers in "Human Intelligence Tasks" (HITs). Whether in the form of fully automated "bots" or as semi-automated tools for completing tasks, automation and Natural Language Generation (NLG) could significantly impact the validity of subjective human insight data collected on crowdsourcing platforms. There is a clear need to discuss, monitor, and study illicit use of automation technology on crowdsourcing platforms.

In mid-2018, a blog post by Bai (2018) on the quality of MTurk data created some attention in news media and academia. Bai (2018) reported "bot-like" responses from crowd workers with similar geographic coordinates and suspected fraudulent use of technology on MTurk. Bai's concerns were picked up by news media (e.g., Dreyfuss (2018); Stokel-Walker (2018)) and received some attention from other researchers in blog posts (Ryan 2018; Moss 2018; Litman 2018) and a limited set of scientific publications (e.g., Chmielewski and Kucker (2020)).

Yet, the research field of crowdsourcing and human computation has paid little attention to this emerging issue. Since the source-agnostic conclusions from Moss (2018) and Litman (2018) three years ago, interest among researchers into investigating the use of automation tools and NLG on crowdsourcing platforms seemingly has ebbed off.

Possible Reasons for the Low Interest Among Researchers and Practitioners in the Field of Crowdsourcing and Human Computation

We speculate and enumerate possible reasons for the low interest in studying the workers' potential use of natural language generation tools on crowdsourcing platforms:

- **Difficulty of the task:** Detecting synthetically generated texts is a hard problem. Humans may be deceived by language models, and cases have been reported in the media of how texts generated with GPT-3 successfully made readers believe that the texts originated from human writers (Hao 2020). While output detector models exist (Gehrmann, Strobelt, and Rush 2019; Solaiman et al. 2019), their accuracy ranges between 72% and 88%. Output detectors additionally are affected by the short length of text collected in microtasks on crowdsourcing platforms which compounds the problem of successfully detecting synthetic texts. Further, research in this space is likely to contribute to a "race for arms" and a "game of cat and mouse" between malicious crowd workers and requesters (Solaiman et al. 2019). The investigation of natural language generation on crowdsourcing platforms is

therefore riddled with many challenges, and may simply be too difficult of a problem for researchers to tackle.

- **Size of the task:** An investigation of natural language generation on crowdsourcing platforms will likely require a large data set. The financial resources needed for such a data collection campaign may exceed the financial resources of a single institution.
- **Unawareness:** MTurk is perused by requesters from different disciplines, including non-technical disciplines. Researchers may simply be oblivious of the possibility for using automated tools on crowdsourcing platforms.
- **Trust in the platform:** Prolific[1] (a crowdsourcing platform for recruiting study participants) has stated that there is "[no] evidence of [...] bot-like accounts" on their platform and that "several processes [are] in place to prevent these types of accounts" (Bradley 2018). Requesters may regard such statements as signals for placing trust in the platform providers to sort out the problem (if there is one).
- **Platform policies and privacy concerns:** The Acceptable Use Policy of MTurk prohibits the collection of "personally identifiable information" (Amazon Mechanical Turk 2018). However, most of the approaches for detecting automation rely on the IP address for investigating the co-location of workers (e.g., (Bai 2018; Ryan 2018; Ahler, Roush, and Sood 2019; Moss 2018; Dennis, Goodson, and Pearson 2020)). While this may be standard practice for many researchers (and part of the default set of metadata collected on survey software, such as Qualtrics), it is a violation of MTurk's Acceptable Use Policy.
- **"End of an era" effect:** Researchers in the field of crowdsourcing may already be aware of issues related to data quality. Researchers may be refraining from problematizing and investigating this issue, because it could be disruptive to their own field of research and the way they conduct their research.
- **"The calm before the storm":** Researchers may be quietly sitting back, waiting for other researchers to take a lead. Another strategy may be to consistently gather data to monitor and accumulate enough data on the issue. This longitudinal approach would require time to become published in the scientific literature.
- **Inertia of scientific publishing:** Investigations of the mid-2018 incident (and other related studies) did not make it through peer review, yet. For instance, the paper by Dennis, Goodson, and Pearson (2020) on the use of Virtual Private Servers (VPS) took almost three years to be accepted in a journal.

We acknowledge that this may not be a complete list, and while the authors are familiar with the literature on crowdsourcing, we may not be aware of all studies on the subject.

Survey of Requesters

In this section, we present our ongoing efforts of clarifying why researchers are not more interested in studying automation, as well as identifying concerns among researchers in crowdsourcing about the use of technology for answering tasks on crowdsourcing platforms. We also inquire about the future of crowdsourcing and human computation in light of the technological developments in ML and NLG.

Survey Design and Participant Recruitment

We created a survey on Google Forms targeting requesters:

https://forms.gle/wznMxVavCmA4z3sN8

Participation is anonymous and the survey does not store personal data (besides basic demographics). The survey consists of 14 items of which two items are open-ended: *"What do you think about automated answer generation (semi-automatic or fully-automatic with 'bots') on crowdsourcing platforms?"* and *"Where do you see the future of crowdsourcing in light of advances in machine learning and the increased availability of tools for natural language generation?"* We also inquire whether researchers are currently or should be in the future concerned about answers being generated on crowdsourcing platforms (on a five-point anchored Likert scale from 'Not At All' to 'Extremely'). The survey includes items to judge the crowdsourcing expertise of the participant (on a scale that is familiar to many participants from peer review systems, such as PCS[2]).

Reaching requesters is a challenge, especially since many researchers are busy and still affected by COVID-19 measures in their countries. We decided on an "opt-in" approach as a gentle and politically correct way of recruiting participants. The call for participation was published in online communities related to crowdsourcing and human computation. We will next publish it in relevant mailing lists.

Preliminary Survey Results

As the study is ongoing and researchers are still invited to participate, we can only report preliminary results. The limited number of responses ($N = 6$; four experts, two with passing knowledge) mentioned that automated answers could lead to bias and cause harm, and that automated answer generation needs to be "detected as much as possible." The participants were moderately to extremely concerned about artificial answer generation. About potential reasons why this topic has not found its way into the scientific literature, the two leading answers were unawareness of researchers and inertia in publishing.

Conclusion

Methods and tools for generating natural-sounding answers to crowdsourcing tasks are bound to find their way into the toolkit of "malicious" crowd workers on crowdsourcing platforms. The technological advances in the area of natural language generation highlight the need for: analyzing and monitoring the extent of the use of automated means for answering questions on crowdsourcing platforms, and the development of reliable measures for collecting and analyzing this information for a good purpose.

[1] www.prolific.co

[2] https://new.precisionconference.com

References

Ahler, D. J.; Roush, C. E.; and Sood, G. 2019. The microtask market for lemons: Data quality on Amazon's Mechanical Turk. In *Meeting of the Midwest Political Science Association*. Midwest Political Science Association.

Amazon Mechanical Turk. 2018. Acceptable Use Policy. URL https://mturk.com/acceptable-use-policy.

Bai, H. 2018. Evidence that A Large Amount of Low Quality Responses on MTurk Can Be Detected with Repeated GPS Coordinates. URL https://www.maxhuibai.com/blog/evidence-that-responses-from-repeating-gps-are-random.

Bradley, P. 2018. Bots and data quality on crowdsourcing platforms. URL https://blog.prolific.co/bots-and-data-quality-on-crowdsourcing-platforms/.

Chmielewski, M.; and Kucker, S. C. 2020. An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results. *Social Psychological and Personality Science* 11(4): 464–473. doi:10.1177/1948550619875149.

Dennis, S. A.; Goodson, B. M.; and Pearson, C. 2020. Online Worker Fraud and Evolving Threats to the Integrity of MTurk Data: A Discussion of Virtual Private Servers and the Limitations of IP-Based Screening Procedures. *Behavioral Research in Accounting* 32(1): 119–134. doi:10.2308/bria-18-044.

Difallah, D. E.; Demartini, G.; and Cudré-Mauroux, P. 2012. Mechanical Cheat: Spamming Schemes and Adversarial Techniques on Crowdsourcing Platforms. In *Proceedings of the 1st International Workshop on Crowdsourcing Web Search (CrowdSearch 2012)*, 26–30.

Dreyfuss, E. 2018. A bot panic hits Amazon's Mechanical Turk. *Wired* URL https://www.wired.com/story/amazon-mechanical-turk-bot-panic/.

Gadiraju, U.; Kawase, R.; Dietze, S.; and Demartini, G. 2015. Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, 1631–1640. New York, NY, USA: ACM. doi:10.1145/2702123.2702443.

Gehrmann, S.; Strobelt, H.; and Rush, A. 2019. GLTR: Statistical Detection and Visualization of Generated Text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 111–116. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/P19-3019.

Hao, K. 2020. A college kid's fake, AI-generated blog fooled tens of thousands. This is how he made it. URL https://www.technologyreview.com/2020/08/14/1006780/ai-gpt-3-fake-blog-reached-top-of-hacker-news/.

Litman, L. 2018. Moving Beyond Bots: MTurk as a Source of High Quality Data. URL https://www.cloudresearch.com/resources/blog/moving-beyond-bots-mturk-as-a-source-of-high-quality-data/.

Moss, A. 2018. After the Bot Scare: Understanding What's Been Happening with Data Collection on MTurk and How to Stop it. URL https://www.cloudresearch.com/resources/blog/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it/.

Ryan, T. J. 2018. Data contamination on MTurk. URL https://timryan.web.unc.edu/2018/08/12/data-contamination-on-mturk/.

Solaiman, I.; Brundage, M.; Clark, J.; Askell, A.; Herbert-Voss, A.; Wu, J.; Radford, A.; Krueger, G.; Kim, J. W.; Kreps, S.; McCain, M.; Newhouse, A.; Blazakis, J.; McGuffie, K.; and Wang, J. 2019. Release Strategies and the Social Impacts of Language Models. URL https://arxiv.org/abs/1908.09203. ArXiv pre-print 1908.09203.

Stokel-Walker, C. 2018. Bots on Amazon's Mechanical Turk are ruining psychology studies. *New Scientist* URL https://www.newscientist.com/article/2176436-bots-on-amazons-mechanical-turk-are-ruining-psychology-studies/.