Outsider Perspectives: Crowd-Based Feedback for Writing

Rhema Linder
Interface Ecology Lab
Texas A&M University
College Station, TX, USA
rhema@ecologylab.net

Shamsi T. Iqbal
Microsoft Research
Redmond, WA, USA.
shamsi@microsoft.com

Jaime Teevan
Microsoft Research
Redmond, WA, USA.
teevan@microsoft.com

CHI'18 Extended Abstracts, April 21–26, 2018, Montréal, QC, Canada.
ACM ISBN 978-1-4503-5621-3/18/04.
http://dx.doi.org/10.1145/3170427.3188602

Abstract
It can be hard for authors to know if what they write will be clear to their readers. While collaborators can provide expert feedback, their limited time and attention makes it costly for authors to continuously solicit detailed input from them. Via a study with ten graduate student authors, we find a clear need for more feedback. Our crowd-based approaches provide an outsider perspective that is timely and detailed, supplementing expert feedback.

Author Keywords
Writing, crowdsourcing, feedback.

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous

Background
Writing involves converting internal thoughts into words that other people can understand, and many common writing errors occur during this translation [5]. Even strong writers are subject to the "curse of knowledge" [17] or "writer-based prose" [5] in which deep familiarity with a context or domain blinds the writer from seeing the text from a reader's perspective. At the moment of writing, the author has an intrinsic sense of what they intend to say. However, predicting

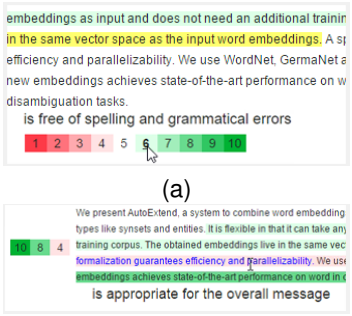

Figure 1: The assess feedback approach, as seen from the perspective of a crowd worker (a) and the end user (b). The crowd worker is rating the yellow highlighted sentence for grammar. The end user has moused over the red sentence and sees the various scores (10, 8, 4) for argument called out.

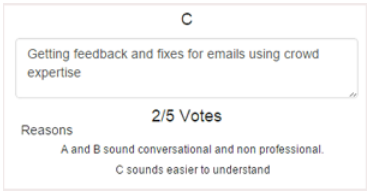

Figure 2: The compare approach as seen by the end user. Option C received 2 out of 5 votes with several reasons.

the clarity of the resulting text and whether it conveys the intended meaning can be difficult.

A common technique authors use overcome this difficulty and improve their writing is to build on other people's feedback [7, 9]. Feedback provides an awareness of the mistakes the author may have made while translating their thoughts into words, and creates the opportunity to fix their writing. Unfortunately, although editors and collaborators typically have the necessary expertise and context to provide useful feedback, their time and attention is limited.

Prior work has looked at how crowd workers, where are more readily available, can support authors in copyediting [1, 10, 20, 2]. However, not as much is currently known about whether feedback provided by crowds, who act as outsiders, are valuable to authors. This goes beyond simple copy editing, condensing, or other simple writing support to focusing on acquiring rich, detailed quality feedback in a timely manner. Even for seemingly simple writing tasks, it takes effort to communicate and represent context from and to crowds [19, 18]. Crowd-based feedback has been used for visual design [8, 11, 12, 15, 1] and student projects [5, 21, 13]. Hicks et al. [11] experimented with crowd-based essay feedback, finding low ratings on essays correlated with comment length. While these studies have shown how crowd-based feedback provides authors with useful insight, they also highlight issues surrounding feedback quality. In other domains, quality issues have been addressed using a rubric, which can help workers focus on salient aspects [6], or by collecting comparative judgments, which have been shown to have lower variance than absolute judgments [3].

Methodology

We explored crowd-based writing feedback by implementing and studying two approaches in the context of academic writing. We begin by detailing the two approaches we used to collect crowd-based writing feedback: *Asses* and *Compare*. To understand the value of these approaches in the context of academic writing, we collected crowd-based feedback on titles and abstracts written by ten graduate student authors (4 male, 6 female), and conducted semi-structured interviews of their reactions. Authors were recruited from the intern pool of a tech company, and compensated with a $10 gift card.

Each of the ten participants provided us with two text items: (1) a set of potential *titles* for a research paper and (2) a self-contained *abstract*. Most provided abstracts that summarized an ongoing research project. On average, the abstracts were just over seven sentences long, and 3 to 5 alternative titles were provided. After the crowd generated feedback, we conducted semi-structured interviews to understand how authors reacted. Interviews were 30 minutes long, and were recorded and transcribed for analysis. First, the authors viewed their feedback with think-aloud. This was followed by authors explaining their reactions, and describing their typical writing tasks and sources of feedback. To analyze, we open coded first, then reorganized around central themes, using a bottom up approach [4].

We collected feedback on the quality of the text produced from crowd workers with a proprietary crowdsourcing platform that outsources to the Clickworker market. The interface is similar to that of Amazon Mechanical Turk; requesters task which workers can choose from a marketplace listing. For each task, we paid US-based workers at least at a rate greater than $10 per hour.Tasks took less than two minutes to complete.

Crowd Assess – Rate by Criteria

To collect assessment feedback on the sentences in a piece of writing, we first extract the sentences. Crowd workers

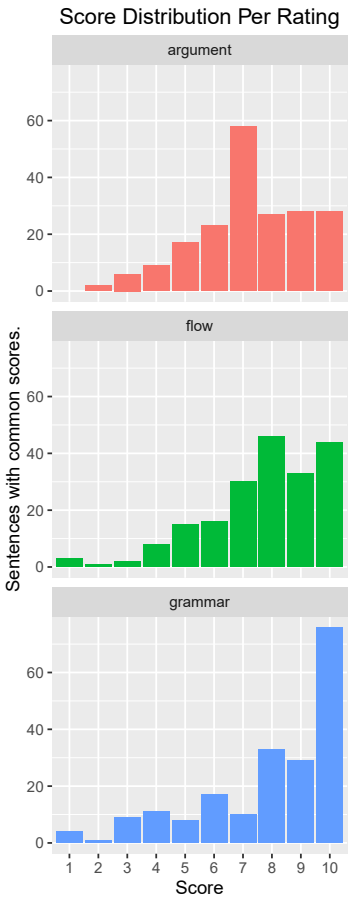

Figure 3: Distribution of crowd-based ratings per sentence per criteria from Assess feedback.

are then asked to rate each sentence on a scale of 1 to 10 based a rubric on one of three criteria: flow "improves the reading of nearby text"), argument ("supports the overall message"), and grammatical correctness ("is free of spelling and grammatical errors"). Because flow and argument require knowledge of the surrounding text, workers are shown the sentence in context of its enclosing paragraph with the sentence highlighted in yellow, as seen in Figure 1(a). The numeric scores are dual encoded with a color, ranging from 1 (negative, red) to 10 (positive, green). Clicking on a score applies that score to the highlighted sentence, sets it to the corresponding color, and advances the task to the next sentence with the same criterion. To leverage context, we ask that workers evaluate all sentences in a given paragraph for a particular criterion at one time, but different workers may assess the paragraph for each criterion. We use multiple workers per criterion to help account for worker variation.

When participants viewed feedback, they were shown an interactive text visualization. High-scoring sentences are shown in green, with low-scoring text in red. Users can filter the writing criterion and mean, max, or min for score summary. Hovering over a sentence, as shown in Figure 1(b), reveals the specific scores from each crowd worker.

Crowd Compare – Pick the Best Option

To collect comparative feedback on a piece of writing from the crowd, workers are shown multiple text alternatives and a single criterion. They are then asked to select which alternative they think is the best alternative and to provide an explanation for their selection.

Authors interested in comparative feedback submit multiple versions of the same text via a web application. Research shows that writing multiple versions of the same text can help authors improve their writing [17]. As the crowd provides feedback, each alternative is annotated with the number of votes it received and corresponding explanations. Figure 2 shows the outcome of Compare as seen by the author.

Figure 3 shows the distribution of all assessment ratings, separated by criteria. While flow has a smooth distribution among scores, argument and grammar have distinct peaks. For argument, scores of 7 occur more than twice as often as than other numbers, perhaps as this criterion was difficult to assess. The distribution for grammar shows a distinct peak at 10. It appears grammar mistakes are mostly binary; errors are either present or they are not. In contrast, the crowd provided more nuanced feedback about argument and flow.

In terms of the comparison feedback, crowd workers tended to agree with each other, with a popular selection by three or more workers occurring 60% of the time. The average length of the reason crowd workers provided for their selection was 10.2 words, ranging from, "Flowed better," to longer, more nuanced explanations.

Results

Our findings show that our participants want more feedback than they currently get from advisors and peers, and that the crowd gave them a valuable outsider perspective that was more timely and detailed than those sources provide.

Need not Met by Current Feedback Sources

Based on our interviews, existing feedback sources do not fully meet our participants' needs for feedback on their academic writing. While all participants reported that there were people available for them to solicit feedback from, the sources of feedback varied. Participants mostly requested feedback from people they worked with closely, including advisors, peers, friends, and partners. While advisors were

Q1.P10: When I finish my first draft, I go through it myself the first time. In the second phase, I give in to my friends who are in my class. . . [Next], I send it to.. my advisor, and he's the one who makes changes and he gets back to me.

Q2.P5: Just sharing the document with things I have written with my co-authors or other people . Having them, usually whenever I send it to them... With my advisor, I sit with them and they go through it and give me feedback as we go. With others [it's] track-changes.. [or] handwriting.

Q3.P7: I would question the clout. . . if I were shooting to publish for SIGGRAPH and I could get people from the SIGGRAPH audience. . . [even] just people who are familiar with that territory [then I would trust the feedback more.]

Q4.P4: Maybe the advisors or people are kind of biased because he really knows the topic and already knows what you talking about.

the most commonly reported source, participants often solicited input from different sources at different stages in the process. For example, several participants reported asking for feedback from their advisor only after having had others look at their writing (Q1.P10).

Typically, the feedback participants received was infrequent and high-level. Advisors and co-authors would annotate text and highlight mistakes (Q2.P5). This concurs with prior work that shows more experienced people, such as advisors, provide smaller, more focused feedback [11].

When feedback participants receive is high-level, it could be hard for them to interpret. Participants complained about advisor comments like, "Poor word choice," or, "Missing key details." P1 explained that her advisor mostly marked for typos, but also noted things that "Looked weird." It was clear from our interviews that participants want more detailed feedback, but did not feel it was consistently available.

Crowd Provides an Outsider Perspective

Participants perceived feedback from the crowd as qualitatively different as they came from non-traditional sources. Not surprisingly, participants had some concerns about the abilities of non-expert crowd workers to provide feedback on academic writing. For example, P7 suggested that he would trust the quality of the feedback more if the pool of workers were more similar to his target audience (Q3.P7).

The crowd workers' lack of in-depth knowledge, however, also seemed to carry with it an opportunity for them to gain an outsider's perspective. Participants, such as P4, felt that external feedback from people not familiar with the research topic were less biased. For example, one participant said that while advisors and co-authors have backgrounds that make it easy for them to understand jargon, the reader of a text may not be as familiar and become lost. Having crowd workers act as a proxy for general readers may help them ensure the text is accessible to those with less contextual knowledge: Likewise, attention-grabbing titles may be hard for someone already familiar with the research to identify. One participant talked about challenges in identifying a good title, and expressed surprise over the one the crowd selected (Q5.P2).

In responding to feedback, participants seemed more comfortable disagreeing with crowd workers than they might be with their advisors or peers. P10, for example, found the lowest-rated feedback to be accurate, but was not concerned with a 4 out of 10 rating in the argument criterion (Q6.P10). Overall, the crowd's outsider perspective appeared to be valuable to participants, providing insight that was not biased by familiarity and that could be acted on without compulsion.

Crowd Feedback Is Timely, Detailed, and Interpreted

Additionally, our participants found the feedback from the crowd was timely and surprisingly detailed. The assessment visualization provided summaries across argument, flow, and grammar criteria. In general, participants focused on flow and argument criteria. For example, P3 describes low flow scores as a "clue." P6 explains that she, "Never gets this kind of feedback," in terms of fine-grained detail (Q7.P6).

For sentences with mixed scores, containing both high and low values, the feedback was still seen as valuable. P5, for example, decided to "go back and understand why" opinions among crowd workers differed. As P5 notes, feedback scores with both high and low scores may indicate a real problem that authors should address.

While sentence-level feedback was appreciated, participants expressed some frustration at not understanding

Q5.P2: My last paper. . . we talked a lot about a title then we came up with five different ones and we sent them to the group to see what people liked. . . I didn't think that [the crowd] would pick C . . . I thought they were all pretty good.

Q6.P10: I don't agree with this [low to neutral rated sentence], actually... and, I agree with this [red, low rated sentence] this deep red one, but the pink is fine. I agree, but it's a paper and I have to argue my thing.

Q7.P6: In this view, I can compare sentences across different things. The granularity is much higher. I can really think about how I'm writing, if the person really could get what I'm saying. Especially this point of, "overall message" [argument]. . . [flow] shows how I can improve my text by removing what is not so important. . . I never get this kind of feedback.

Q8.P3: It seems that I had, the most actual problems are about the flow, how the sentences are connected. Which makes sense, but, the only thing that I can say is that it doesn't exactly tell me what I should change. But, it gives me a clue.

why some sentences scored low. The most requested feature was for comments, as illustrated by P3's comments (Q8.P3). The observation that low ratings merit explanation aligns with findings that crowd workers that rate an overall essay as poor provide more qualitative feedback [11]. Unlike the Assess approach, the Compare approach does provide explanations, which participants found helpful.

Conclusion

In this paper, we explored the potential value of crowd-based writing feedback by implementing two feedback collection approaches. In the first approach, crowd workers assessed the quality of a paper abstract along dimensions of grammar, argument and flow. In the second, crowd workers compared several title alternatives offered by the author. We studied these approaches by collecting feedback on the titles and abstracts written by ten graduate student authors and interviewing them about their reactions.

Participants felt that the feedback from the crowd was complimentary to the feedback they received from other sources, which are not always available. The crowd-based feedback seemed less biased, more timely, and detailed. While the crowd does not have the expertise that an advisor or peer may have, it appears to offer an easily accessible outsider perspective that can be valuable for people while writing. Participants did not need to agree with the feedback they received, and did not feel compelled to act on all of the feedback they received. Even in these cases, however, it helped them to reconsider their approach or feel confident they made the right decision in parts of their original writing.

These findings suggest that feedback systems can benefit from a range of opinions, including the crowd's. Although participants were able to benefit from crowd feedback they disagreed with, maximizing feedback quality is important. Combining other common crowd-based quality approaches [14] with our rubric and comparison-based approaches may better address this. The best approach may depend on the context in which the feedback is used. For example, although Compare judgments tended to have low variance [3], Assess might still be most appropriate for longer text where it is hard to collect multiple versions. There may also be ways to identify expert or personalized [16] crowds.

REFERENCES

1. Elena Agapie, Jaime Teevan, and Andrés Monroy-Hernández. 2015. Crowdsourcing in the field: A case study using local crowds for event reporting. In *AAAI Human Computation and Crowdsourcing*.
2. Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. Soylent: A Word Processor with a Crowd Inside. In *Proc UIST*. ACM, New York, NY, USA, 313–322. DOI: http://dx.doi.org/10.1145/1866029.1866078
3. Ben Carterette and Paul N. Bennett. 2008. Evaluation Measures for Preference Judgments. In *Proc SIGIR*. ACM, New York, NY, USA, 685–686. DOI: http://dx.doi.org/10.1145/1390334.1390451
4. Juliet Corbin and Anselm Strauss. 2008. Basics of qualitative research: Techniques and procedures for developing grounded theory. (2008).
5. Steven Dow, Elizabeth Gerber, and Audris Wong. 2013. A Pilot Study of Using Crowds in the Classroom. In *Proc CHI*. ACM, New York, NY, USA, 227–236. DOI: http://dx.doi.org/10.1145/2470654.2470686
6. Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the crowd yields better work. In *Proc CSCW*. ACM, 1013–1022.

7. Linda Flower and John R Hayes. 1981. A cognitive process theory of writing. *College composition and communication* 32, 4 (1981), 365–387.
8. Michael D. Greenberg, Matthew W. Easterday, and Elizabeth M. Gerber. 2015. Critiki: A Scaffolded Approach to Gathering Design Feedback from Paid Crowdworkers. In *Proc ACM Creativity and Cognition*. ACM, New York, NY, USA, 235–244. DOI: http://dx.doi.org/10.1145/2757226.2757249
9. Nick Greer, Jaime Teevan, and Shamsi T Iqbal. 2016. *An introduction to technological support for writing*. Technical Report. Technical Report. Microsoft Research Tech Report MSR-TR-2016-001.
10. Nathan Hahn, Joseph Chang, Ji Eun Kim, and Aniket Kittur. 2016. The Knowledge Accelerator: Big Picture Thinking in Small Pieces. In *Proc CHI*. ACM, New York, NY, USA, 2258–2270. DOI: http://dx.doi.org/10.1145/2858036.2858364
11. Catherine M. Hicks, Vineet Pandey, C. Ailie Fraser, and Scott Klemmer. 2016. Framing Feedback: Choosing Review Environment Features That Support High Quality Peer Assessment. In *Proc CHI*. ACM, New York, NY, USA, 458–469. DOI: http://dx.doi.org/10.1145/2858036.2858195
12. Julie Hui, Amos Glenn, Rachel Jue, Elizabeth Gerber, and Steven Dow. 2015. Using anonymity and communal efforts to improve quality of crowdsourced feedback. In *AAAI Human Computation and Crowdsourcing*.
13. Julie S. Hui, Elizabeth M. Gerber, and Steven P. Dow. 2014. Crowd-based Design Activities: Helping Students Connect with Users Online. In *Proc DIS*. ACM, New York, NY, USA, 875–884. DOI: http://dx.doi.org/10.1145/2598510.2598538
14. Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proc CSCW*. ACM, 1301–1318.
15. Kurt Luther, Jari-Lee Tolentino, Wei Wu, Amy Pavel, Brian P. Bailey, Maneesh Agrawala, Björn Hartmann, and Steven P. Dow. 2015. Structuring, Aggregating, and Evaluating Crowdsourced Design Critique. In *Proc CSCW*. ACM, New York, NY, USA, 473–485. DOI: http://dx.doi.org/10.1145/2675133.2675283
16. Peter Organisciak, Jaime Teevan, Susan Dumais, Robert C Miller, and Adam Tauman Kalai. 2014. A crowd of your own: Crowdsourcing for on-demand personalization. In *AAAI Human Computation and Crowdsourcing*.
17. Steven Pinker. 2014. The source of bad writing. *The Wall Street Journal* (2014).
18. Niloufar Salehi, Jaime Teevan, Shamsi T Iqbal, and Ece Kamar. 2017. Communicating Context to the Crowd for Complex Writing Tasks. In *CSCW*. 1890–1901.
19. Carolin Shah, Katharina Erhard, Hanns-Josef Ortheil, Evangelia Kaza, Christof Kessler, and Martin Lotze. 2013. Neural correlates of creative writing: an fMRI study. *Human brain mapping* 34, 5 (2013), 1088–1101.
20. Jaime Teevan, Daniel J Liebling, and Walter S Lasecki. 2014. Selfsourcing personal tasks. In *Proc CHI EA*. ACM, 2527–2532.
21. Helen Wauck, Yu-Chun (Grace) Yen, Wai-Tat Fu, Elizabeth Gerber, Steven P. Dow, and Brian P. Bailey. 2017. From in the Class or in the Wild?: Peers Provide Better Design Feedback Than External Crowds. In *Proc CHI*. ACM, New York, NY, USA, 5580–5591. DOI: http://dx.doi.org/10.1145/3025453.3025477