Evaluating and Complementing Vision-to-Language Technology for People who are Blind with Conversational Crowdsourcing

Elliot Salisbury*, Ece Kamar+, Meredith Ringel Morris+
*University of Southampton, +Microsoft Research
*e.salisbury@ecs.soton.ac.uk, + {eckamar, merrie}@microsoft.com

Abstract

We study how real-time crowdsourcing can be used both for evaluating the value provided by existing automated approaches and for enabling workflows that provide scalable and useful alt text to blind users. We show that the shortcomings of existing AI image captioning systems frequently hinder a user's understanding of an image they cannot see to a degree that even clarifying conversations with sighted assistants cannot correct. Based on analysis of clarifying conversations collected from our studies, we design experiences that can effectively assist users in a scalable way without the need for real-time interaction. Our results provide lessons and guidelines that the designers of future AI captioning systems can use to improve labeling of social media imagery for blind users.

Introduction

As social media is becoming pervasive in American culture [Duggan *et al.*, 2015], it is important that people who are blind or visually impaired (BVI) can access the entirety of content shared in social media. However, embedded imagery is becoming more prevalent in social media; a study of Twitter found that more than 40% of popular (retweeted) posts contained embedded multimedia as of June 2015 [Morris *et al.*, 2016], which constrains the accessibility of the content in Twitter by BVI users. While Twitter recently began to offer limited capabilities to augment images with alternative text (a.k.a. alt text or captions) that can be read aloud by screen reader technology [Kloots, 2016]; alt text compliance and quality on the web in general is low [Bigham *et al.*, 2006; Goodwin *et al.*, 2011]. Recently, automated approaches that combine computer vision and natural language processing to describe image content have emerged as a potential solution for improving the accessibility of social media imagery for BVI users. Examples include the automatic alt text system deployed by Facebook [Wu *et al.*, 2016] and automated image captioning systems [Fang *et al.*, 2015; Karpathy and Fei-Fei, 2015]. Although assisting blind users is a motivating application domain for these systems, the value these imperfect systems provide to BVI users is unclear. While existing systems are tested in the lab within constrained data sets, the performance of these systems in the context of social media (which incorporates a wide variety of professional and casual quality imagery and covers a range of subjects and styles) is not yet studied. Unexpected imperfections in automated system output may degrade user trust, or may negatively impact users instead of helping them.

In this work, we explore ways for combining crowd input and automated approaches to assist BVI users in accessing social media with visual content. Our studies focus on the following questions: (1) What value is provided by a state-of-the-art vision-to-language API in assisting BVI users, and what are the areas for improvement? (2) What are the trade-offs between alternative workflows for the crowd assisting BVI users? (3) Can human-in-the-loop workflows result in reusable content that can be shared with other BVI users?

To study these research questions, we designed and experimented with workflows that varied the level of human engagement and the involvement of an automated system to better understand the requirements for creating good-quality, scalable, automated or semi-automated alt text for BVI consumers of social media. The results show that the negative impact of erroneous system output on user understanding is so significant that it cannot be completely erased even through free-form conversation with a sighted assistant. On the positive side, human input, either assisting users alone or correcting/complementing the automated system, is effective in increasing user satisfaction. Our structured Q&A workflow is shown to be effective for enabling scalable, lower-cost assistance to BVI users. We complement the large-scale crowdsourcing study with a small-scale evaluation of TweetTalk with real BVI users. We conclude with a set of guidelines that future work can use to improve labeling of social media imagery for blind users.

This extended abstract provides a summary of work published in [Salisbury *et al.*, 2017]. Please see the longer paper for examples, and details on experiements and results.

Related Work

Crowdsourced conversational interfaces have been developed for assisting BVI users with their daily tasks [Bigham *et al.*, 2010; Lasecki *et al.*, 2013]. Social Microvolunteering [Brady *et al.*, 2015] uses third-party friendsourcing to achieve low-cost, high-quality answers to visual questions from people

who are BVI, but it is unclear that the technique is scalable to provide alt text to large sets of online images.

A recent study by [MacLeod *et al.*, 2017] investigated how BVI users perceive captions generated by automated approaches for a curated set of image tweets. MacLeod, et al. showed that BVI users trust auto-generated captions even when they are inaccurate, and studied how to convey skepticism to prevent over-trusting. In this paper, we focus on human-in-the-loop workflows to improve the value BVI users get from alt text.

Workflows for Alt Text Generation

We designed and studied four workflows for providing an understanding of images accompanying tweets, with BVI users as the target audience. The inputs to each workflow are a single tweet, containing the tweet's text and the accompanying image. Then each workflow attempts to explain the tweet's image to BVI users within the context of the tweet.

The first two workflows provide a baseline state of the art approach to captioning images. The first workflow, Vision-to-Language, uses captions generated by the CaptionBot system for the tweet's image [Fang *et al.*, 2015]. CaptionBot is based off the technology that won the 2015 CVPR captioning challenge, and uses Microsoft Cognitive Services [Microsoft, 2016], a set of APIs used for understanding imagery and text. Current Vision-to-Language systems cannot yet use the additional context of the tweet text and purely caption the image instead. The second workflow, Human-Corrected Captions, provides crowd workers with the original tweet text, accompanying image, and the Vision-to-Language-generated alt text. While human corrections may fix factual errors in the automatically generated captions, the value of human-corrected captions to BVI users may be limited since workers may not foresee the type of information or the level of detail required for high-quality alt text desired by BVI users.

We developed two subsequent experiences, the TweetTalk conversational assistant workflow and the Structured Q&A workflow, that build upon and enhance the baseline captions. These four workflows allow us to investigate what key information end users desire in a caption for a social media image, how effective deeper human assistance is, and whether the information desired by a single consumer of the alt text will satisfy a larger audience of end users.

The workflows were tested on Amazon's Mechanical Turk (AMT), with recruitment restricted to U.S. workers only, due to the collection of tweets being mainly U.S.-centric and the description and conversations these workers take part in requiring sufficient understanding of the English language. Because current crowdsourcing platforms are largely inaccessible to users with disabilities and therefore lack a sufficiently large pool of BVI workers [Zyskowski *et al.*, 2015], when testing our workflows, we simulated the experience of being BVI by employing (presumably) sighted turkers (whom we will refer to as the simulated-BVI workers) and making the images unavailable to them. While necessitated by practical constraints of testing these workflows at scale, we recognize that simulated-BVI workers may have different captioning preferences than people who are BVI; hence, we conducted additional testing with seven people who are blind or visually impaired to validate the generalizability of our findings.

We experiment with a data set of 85 tweets that was curated by previous work [MacLeod *et al.*, 2017]. Each tweet contains embedded image attachments from a set of popular accounts (e.g., @HillaryClinton, @nytimes, @TaylorSwift) and/or trending hashtags (e.g., #tbt [throw-back Thursday]). Tweets were selected to cover a broad range of topics (e.g., humor, news, celebrities, memes, etc.), representing the varied interests reported by blind users of Twitter [Morris *et al.*, 2016]. The tweets vary in terms of confidence of the automated system in generating an auto-caption.

Conversational Assistant Workflow

The Conversational Assistant workflow uses TweetTalk, a scalable conversational platform between BVI (or simulated-BVI) users and human assistants. TweetTalk allows BVI users to have free-form conversations with sighted workers to find out about visual content. Analyses of conversations collected from TweetTalk show us what kind of information BVI users are interested in, extract key classes of information that can help enhance captions, and measure the value gained from unconstrained human assistance.

Our conversational assistant platform, TweetTalk, was built on top of the architecture described in [Mao *et al.*, 2012], and enables us to investigate conversations between sighted and simulated-BVI crowd workers about a given tweet. This workflow connects two workers, but provides each worker with a different interface. One worker, whom we will refer to as the sighted assistant, can see the imagery associated with the tweet, while the other (the simulated-BVI worker) cannot. The simulated-BVI worker must then have a conversation with the sighted assistant in order to understand the image accompanying the tweet and write a description of it.

The workflow employs the following steps (See Figure 1):

1. **Read the tweet:** Both workers are shown the tweet's text and author and a Baseline Image Caption, that could either be empty, generated from Vision-to-Language, or a Human-Corrected caption. This baseline caption seeds the simulated-BVI worker's understanding of the image. Only the sighted assistant is shown the image associated with the tweet.

2. **Rate the caption:** We ask only the simulated-BVI worker to rate the utility of the baseline caption (if there is one), as they have not yet seen the image, so we can assess the initial trust the BVI user has for the baseline caption, and how this assessment later changes as a result of gaining more information about the image through the following conversation.

3. **Ask/Answer questions:** Both workers have access to a chat box; the simulated-BVI worker is asked to initiate the conversation by asking one or more questions about the image. The sighted assistant is asked to reply sensibly to these questions, but without writing their own complete description of the image, because we are interested in capturing the simulated-BVI worker's questions. This step has two purposes; it informs us about the

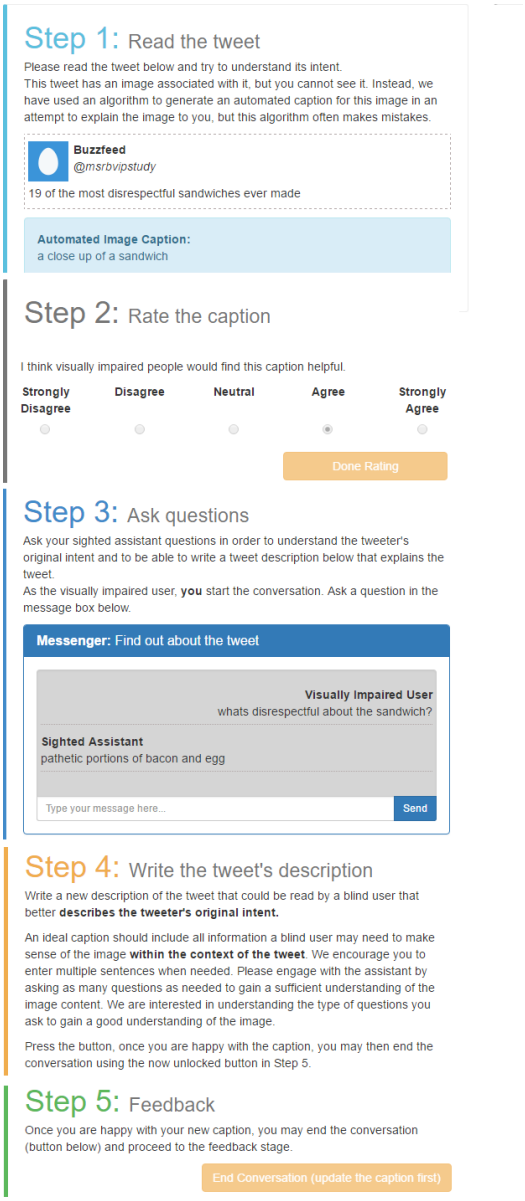

Figure 1: An example of the TweetTalk interface shown to workers in the simulated-BVI role.

information users would like, and allows us to quantify the effectiveness of free-form human assistance.

4. **Write a description:** After they terminate the interaction with the sighted assistant, the simulated-BVI worker is then asked to write a new description of the tweet's image, so that we can evaluate their understanding gained through conversation.
5. **Feedback:** In the final step, we show the simulated-BVI worker the image for the first time, and ask them to rerate the baseline caption and the new description generated in step 4. As such, we can gain insight into their assessment of the effectiveness of the conversation.

We tested the system by recruiting 235 unique workers for the conversational assistant. To rate the baseline image captions and the simulated-BVI worker's generated descriptions, workers are asked the same Likert-type question (i.e., "I think visually impaired people would find this caption helpful.") used in [MacLeod *et al.*, 2017], using a five-point scale (Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree). During the conversational workflow, the simulated-BVI workers taking part in the conversations are asked to rate the baseline caption twice, once before the conversation having not seen the image (First-Party Before for Baseline Captions), and again after the conversation once the image is revealed to them (First-Party After for Baseline Captions). By asking the same question before and after seeing the image, we can then measure any change in ratings after seeing the image, which we refer to as the satisfaction factor. This factor captures how misleading a caption may be to a BVI user.

Once the simulated-BVI worker finishes conversing with the sighted assistant, the simulated-BVI worker is asked to write a description of the image to the best of their understanding, gained through conversing with the sighted assistant without seeing the image. This allows us to extract how well they have understood the imagery. Once the image is revealed to the simulated-BVI user, they then must rate their description (First-Party After for Descriptions Generated Through TweetTalk Conversations).

Structured Questions Workflow

We developed a streamlined workflow using the most common question types present in TweetTalk interactions. This workflow eliminates the need for a conversational pairing, thus reducing the temporal and monetary costs, and removing the common conversational problem of user dropouts.

Conversational Analysis

We analyzed the 429 questions that were sent by simulated-BVI workers when conversations were seeded by the Human Corrected Captions using an iterative, open coding approach in which we identified and refined thematic categories in the questions. From these coded questions, we identified the core concepts that our users were interested in. We used these concepts to create a set of questions to extract desired details about social media images. Examples of questions included in the list include: "Who are the main subjects of the image? Describe their physical characteristics.", "Describe the location and the prominent features of the background", "What are the subjects of the image doing?", "What emotion does this image evoke?", "Is this intended to be humorous? Explain how.", "Is this a famous or well-known image?", and "Describe noteworthy aspects of this image's visual style."

To evaluate the effectiveness of these answers to understanding the image, we adopt a similar interface as that shown to the simulated-BVI worker in TweetTalk, replacing the chat box with the list of answered questions and, as before, the worker is asked to write a description of the tweet.

Experiments

We evaluated the Conversational Assistant workflow with 235 unique crowd workers. We held one conversation per tweet

	Baseline Image Captions		Descriptions Generated Through TweetTalk Conversations			
	Vision-to-Language	Human-Corrected Captions	No Caption	Vision-to-Language	Human-Corrected Captions	Structured Questions
First-Party Before	2.56	3.36				
First-Party After	1.92	3.48	4.11	3.97	4.22	4.11
First-Party Satisfaction	-0.64	0.12				
Third-Party Before	2.91	3.63	3.70	3.92	3.81	3.42
Third-Party After	1.85	3.74	3.65	3.64	3.83	4.10
Third-Party Satisfaction	-1.06	0.11	-0.05	-0.28	0.02	0.68

Table 1: Average Likert Ratings: Before/After ratings are on a 1-5 scale; Satisfaction ratings are on a -4 to 4 scale (i.e., how much an individual changes their rating after seeing the image); higher ratings are better.

per initial seeding Baseline Image Caption, and with no seeding captions at all (i.e., 3 treatments), leading to a total of 255 conversations. Evaluating the Structured Questions workflow does not require worker pairing and thus was faster and easier to run; we performed three repeats per tweet, for a total of 255 runs. In addition to first-party evaluations, the baseline image captions and the descriptions generated through our workflows were also evaluated by crowd workers who did not participate in the conversation (Third-party Evaluation).

Table 3 presents the first-party and third-party ratings of both the simple baseline image captions, the descriptions generated through the Conversational Assistant workflow (for each treatment seeding the conversation with a different caption), and the Structured Questions workflow. All results stated as significant have been found as such using Friedman's test with a follow-up pairwise comparison using Wilcoxon's test with Bonferroni correction.

We found no significant difference across conditions for the first-party ratings (i.e., the rating they give their own description after the task). We also observed that the first-party ratings were higher than those given by third-parties uninvolved in their creation. To further investigate this disparity, we designed a quick follow up study in which we showed both the description and the conversation to third-party raters and checked if this disparity was due to intrinsic valued gained through conversation that wasn't relayed in their description. The results suggested that showing the conversation did not provide additional value and the disparity results from workers rating their own work higher.

Next, we evaluate the captions and descriptions generated by various workflows based on their third-party evaluations (i.e., collected after seeing the imagery). The results show that current Vision-to-Language systems have significantly worse accuracy when compared to even a simple human-in-the-loop approach (Human-Corrected Captions, z=7.45, $p < .001$) and to our caption improvement workflows (Conversational Assistant and Structured Questions, z=2.20, $p < .001$ and z=3.19, $p < .001$). This suggests that automatic image captioning systems require more work before they are ready for use by social networking platforms.

We observe no significant difference in the accuracy between the Human-Corrected Captions and the description generated after using TweetTalk, on the treatments seeded with either no caption or the Human-Corrected captions (those seeded with Vision-to-Language captions are discussed below). However, the Structured Questions approach significantly ($0.52 <= z <= 0.99$, $p <= .03$) improves understanding against all approaches.

Additionally, we observed that seeding the conversation with Vision-to-Language creates significantly less satisfaction, (i.e., the captions are believable, but turn out to be inaccurate) than simply providing a Human-Corrected Caption ($z = -0.78, p < .001$), or conversations seeded with Human-Corrected Captions ($z = -0.63$, $p = .003$).

Another consideration for the comparison of workflows is the time and monetary costs of assisting BVI users. Real-time crowdsourcing for free-form conversation is time consuming and expensive, on average taking 8 minutes per tweet, and costing up to $0.95 for compensating the sighted assistant. Whereas, the structured questions do not suffer from the challenges of real-time crowdsourced conversation; the time taken to answer a question on average takes 1 minute, and although we need multiple workers to answer the same question, these can be performed simultaneously. The total cost of these HITs, to get 3 answers to the 8 questions, was $1.20. Although more expensive than the human-corrected captions and the conversational assistant, the structured Q&A workflow is more general purpose, results in a much greater satisfaction, and the cost can be amortized across multiple BVI users, while the conversation is an individual experience. In future versions of the of the structured Q&A workflow, different strategies such as answering multiple questions per HIT, or predicting relevant questions to ask per tweet, can be taken to reduce its cost.

Validation with BVI Users

We ran a follow-up study with seven blind adults. Given the limited size of our subject pool, we preferred to use TweetTalk over Structured Q&A in experimenting with real BVI users so that we could collect more detailed information than just assessments, including what BVI users ask about and their preferences about interactive crowd experiences.

For these experiments, the the role of the sighted assistant was fulfilled by a member of our research team due to screen reader accessibility issues.

The questions the BVI users asked were coded using the same scheme as before; no new types of questions were asked, and there was no significant difference in the frequency of these question types, indicating that the Structured Q&A workflow would be informative for real BVI users.

Conclusions and Future Work

We have shown how current AI captioning systems may hinder, rather than help, BVI users' understanding of social media posts. We developed workflows that incorporate different levels of automation and human involvement to improve this understanding, and to analyze the information that BVI users wish to know about.

There is value in exploring alternative alt text formats, such as interactive formats in which users can query additional information about the image should they wish, perhaps using our set of structured questions. For popular imagery and posts, the guideline questions could be pre-asked, anticipating the details users would want. For those questions not yet asked before, real-time crowdsourcing could be used to respond quickly, and any future similar question can return the same answer, reducing the workload and distributing the cost.

References

[Bigham *et al.*, 2006] Jeffrey P Bigham, Ryan S Kaminsky, Richard E Ladner, Oscar M Danielsson, and Gordon L Hempton. Webinsight:: making web images accessible. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, pages 181–188. ACM, 2006.

[Bigham *et al.*, 2010] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342. ACM, 2010.

[Brady *et al.*, 2015] Erin Brady, Meredith Ringel Morris, and Jeffrey P Bigham. Gauging receptiveness to social microvolunteering. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1055–1064. ACM, 2015.

[Duggan *et al.*, 2015] Maeve Duggan, Nicole B Ellison, Cliff Lampe, Amanda Lenhart, and Mary Madden. Social media update 2014. pew research center, 2015.

[Fang *et al.*, 2015] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1473–1482, 2015.

[Goodwin *et al.*, 2011] Morten Goodwin, Deniz Susar, Annika Nietzio, Mikael Snaprud, and Christian S Jensen. Global web accessibility analysis of national government portals and ministry web sites. *Journal of Information Technology & Politics*, 8(1):41–67, 2011.

[Karpathy and Fei-Fei, 2015] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

[Kloots, 2016] Todd Kloots. Accessible images for everyone. `https://blog.twitter.com/2016/accessible-images-for-everyone`, 2016.

[Lasecki *et al.*, 2013] Walter S Lasecki, Rachel Wesley, Jeffrey Nichols, Anand Kulkarni, James F Allen, and Jeffrey P Bigham. Chorus: a crowd-powered conversational assistant. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pages 151–162. ACM, 2013.

[MacLeod *et al.*, 2017] Haley MacLeod, Cynthia L. Bennett, Meredith Ringel Morris, and Edward Cutrell. Understanding blind peoples experiences with computer-generated captions of social media images. In *Proceedings of the 2017 CHI Conference*, 2017.

[Mao *et al.*, 2012] Andrew Mao, Yiling Chen, Krzysztof Z Gajos, David Parkes, Ariel Procaccia, and Haoqi Zhang. Turkserver: Enabling synchronous and longitudinal online experiments. In *Proceedings of the Fourth Workshop on Human Computation (HCOMP'12)*. AAAI Press, 2012.

[Microsoft, 2016] Microsoft. Microsoft cognitive services api. `https://www.microsoft.com/cognitive-services/en-us/computer-vision-api`, 2016.

[Morris *et al.*, 2016] Meredith Ringel Morris, Annuska Zolyomi, Catherine Yao, Sina Bahram, Jeffrey P Bigham, and Shaun K Kane. With most of it being pictures now, i rarely use it: Understanding twitter's evolving accessibility to blind users. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5506–5516. ACM, 2016.

[Salisbury *et al.*, 2017] Elliot Salisbury, Ece Kamar, and Meredith Ringel Morris. Toward scalable social alt text: Conversational crowdsourcing as a tool for refining vision-to-language technology for the blind. In *Proceedings of the fifth AAAI Conference on Human Computation and Crowdsourcing*, 2017.

[Wu *et al.*, 2016] S Wu, H Pique, and J Wieland. Using artificial intelligence to help blind people see facebook. `http://newsroom.fb.com/news/2016/04/using-artificial-intelligence-to-help-blind-people-see-facebook/`, 2016.

[Zyskowski *et al.*, 2015] Kathryn Zyskowski, Meredith Ringel Morris, Jeffrey P. Bigham, Mary L. Gray, and Shaun Kane. Accessible crowdwork? understanding the value in and challenge of microtask employment for people with disabilities. ACM Association for Computing Machinery, March 2015.