Introducing Game Elements in Crowdsourced Video Captioning by Non-Experts

Hernisa Kacorri
The Graduate Center, CUNY
365 Fifth Ave, New York,
NY 10016 USA
hkacorri@gc.cuny.edu

Kaoru Shinkawa
IBM Research – Tokyo
5-6-52 Toyosu, Koto-ku
Tokyo, 135-8511 Japan
kaoruma@jp.ibm.com

Shin Saito
IBM Research – Tokyo
5-6-52 Toyosu, Koto-ku
Tokyo, 135-8511 Japan
shinsa@jp.ibm.com

ABSTRACT

Video captioning can increase the accessibility of information for people who are deaf or hard-of-hearing and benefit second language learners and reading-deficient students. We propose a caption editing system that harvests crowdsourced work for the useful task of video captioning. To make the task an engaging activity, its interface incorporates game-like elements. Non-expert users submit their transcriptions for short video segments against a countdown timer, either in a "type" or "fix" mode, to score points. Transcriptions from multiple users are aligned and merged to form the final captions. Preliminary results with 42 participants and 578 short video segments show that the Word Error Rate of the merged captions with two users per segment improved from 20.7% in ASR to 16%. Finally, we discuss our work in progress to improve both the accuracy of the collected data and to increase the crowd engagement.

Categories and Subject Descriptors

K.4.2 [**Computer and Society**]: Social issues—Assistive technologies for persons with disabilities; H.5.3 [**Group and Organization Interfaces**]: Collaborative computing.

General Terms

Design, Human Factors.

Keywords

Crowdsourcing, video captioning, transcription, gamification.

1. INTRODUCTION

Video captioning can contribute towards bridging the digital divide for people who are deaf or hard-of-hearing by providing access to news, entertainment, and information. It has universal design benefits, beyond its support for video indexing and summarization. Video captioning and audio transcriptions can improve comprehension, fluency, and literacy skills for people who are learning a second language, e.g. [7][16]. Displaying videos with captions at an average or slow speed can also help students with reading deficiencies, as shown in [6][10]. Captions are also beneficial for people in noisy environments such as airports, subways, and sports clubs or in places where noisy speakers cannot be used.

While Automatic Speech Recognition (ASR) is often used to add captions to video content, as in [2][3][5][15], the results can be poor with real-time recognition, noisy environments, multilingual or multi-accented audio, and informal or untrained speakers. In contrast, including human transcribers in the loop may pose cost and technical challenges, requiring experts in real-time captioning, efficient and language-competent typists, effective user interfaces to minimize errors, and complex quality assurance methods. To solve the computational problems of video captioning, researchers such as [8][9] have investigated splitting the tasks into smaller segments and asking MTurk workers [1] for partial solutions that are later combined into the final results. However, we want non-monetary incentives to motivate crowd engagement and control the costs of captioning.

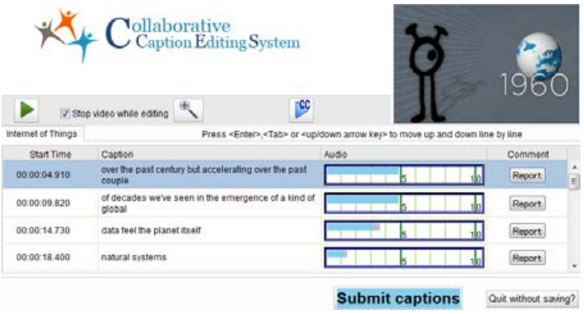
In this paper, we propose a crowdsourcing platform for obtaining video captions from a crowd of non-experts without monetary rewards. We describe the transition from an initial caption-editing tool with expert editors in the loop to a new editing tool merging non-experts' input and incorporating game-like incentives (Section 2). We show how an initial development prototype works (Section 3), analyze preliminary results obtained from a small group of users (Section 4), and discuss our ongoing work to improve the proposed solution (Section 5).

2. COLLABORATIVE CAPTION EDITING

The framework for our proposed work is the Collaborative Caption Editing System (CCES) [12]. Initially, video submitted by a content owner is automatically divided into 30-second to 120-second pieces and distributed for the next step of captioning by many non-expert editors. For better captioning, the system next divides these mid-sized video chunks into linked meaningful segments of speech by detecting phrase boundaries in the audio signal. Each editor listens to these snippets and either types the words in the video segment or edits the ASR text (when such correction is enabled). Figure 1 illustrates a screenshot of the caption-editing tool, where the video display window is shown above the text lines corresponding to the short video segment. For each of the lines available for editing, there is a position-within-segment indicator synchronized with the playing video. The user interface is designed to be intuitive and tuned to make captions in the smallest number of operations. For each line, which is a 2-second to 10-second video segment, the voiced and silent areas are detected. The blue part in the audio ruler is the voiced area, and the gray part is the silent area. When a caption line is in the focus, the input field of that line is enabled, and the corresponding audio and video plays automatically in a loop until the editor

W4A '14, April 07 - 09 2014, Seoul, Republic of Korea
Copyright 2014 ACM 978-1-4503-2651-3/14/04…$15.00.
http://dx.doi.org/10.1145/2596695.2596713

moves to the next caption line. While looping, the silent parts are skipped for higher efficiency. These partial transcriptions are collected from all of the editors and combined to form the captions for the entire video. As a final step, expert editors can verify the captions for the full video and make any needed final corrections.



Press <Enter>,<Tab> or <up/down arrow key> to move up and down line by line

Start Time	Caption	Audio	Comment
00:00:04.910	over the past century but accelerating over the past couple	5 10	Report
00:00:09.820	of decades we've seen in the emergence of a kind of global	5 10	Report
00:00:14.730	data feel the planet itself	5 10	Report
00:00:18.400	natural systems	5 10	Report

Figure 1. Original CCES Editing Tool.

To eliminate the need for experts in the last step of quality verification, we propose a new (CCES) editing tool (Fig. 2) that allows multiple non-expert editors to work on the same video and automatically merge their suggestions into a final result. This interface seeks to engage the editors by incorporating game-like elements. Relatively few researchers have considered such game-like elements as scoring [8][9][15] or opposing teams [9] for crowdsourced captioning. However, many of the proposed video captioning and audio transcribing systems do incorporate monetary rewards [8][9][3][4].

2.1 Game-Like Elements in CCES Editing

Captions for the segmented videos are collected from multiple editors in two modes:

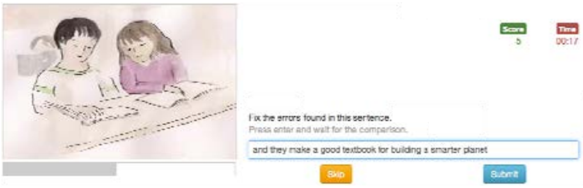
- **Type Mode**: The user listens to the video and types a transcription, or
- **Fix Mode**: The ASR results are used as a suggested transcription that requires further checking and editing.

Each video segment is 2 to 10 seconds long and plays in a loop. However, the clip pauses when the user is editing and resumes when no keyboard input is detected. Fig. 2 shows a screenshot of the proposed editing tool in Fix mode. A bar under the video clip indicates the position of the video when it plays, stops, and resumes. Users can skip a video clip, if they are not sure about the words, or submit their transcriptions.

The experimental interface we studied incorporates two game-like elements: Limited editing time for the video segment (using a countdown clock) and scoring feedback for the submitted transcription.

Countdown clock: When the time expires the user's current transcription is automatically submitted. The time limit is defined by the duration of that particular video clip multiplied by a constant factor. This multiplier controls how challenging the game feels, so the proper value is a key to sustaining engagement meanwhile allowing for adequate editing time. Besides being an essential game-like mechanism, a countdown clock allows us to investigate near real-time captioning with the proposed tool.

Scoring: Since initially we don't know the correct transcript for the video clip, an approximate score is assigned to the user's submitted transcript. In Type mode, this is based on the similarity to the ASR result, while in Fix mode it is compared with another user's transcript. The number of matching words is the score.



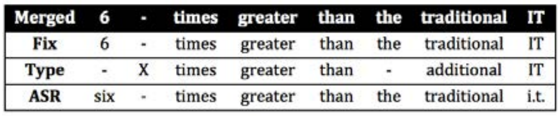
Fix the errors found in this sentence.
Press enter and wait for the comparison.

and they make a good textbook for building a smarter planet

Figure 2. New CCES Editing Tool in Fix Mode.

2.2 Combining Multiple Editors' Input

For each video segment, the submitted transcriptions in both Type and Fix modes are aligned with the text suggested by the speech recognition engine. For merging aligned phrases, majority voting is used, where a word is accepted if it appears at least twice. In the initial CCES editing tool, each video segment was only given to one participant and then checked by an expert. With the new tool, each segment can be given to multiple participants and a merging algorithm is used in lieu of the experts. All the caption fragments resulting from the merge step are automatically combined to form the full captions for the entire video. Figure 3 illustrates an example of alignment and merging with an ASR result, a transcription submitted in Type mode, and a second text submitted in Fix mode.



Merged	6	-	times	greater	than	the	traditional	IT
Fix	6	-	times	greater	than	the	traditional	IT
Type	-	X	times	greater	than	-	additional	IT
ASR	six	-	times	greater	than	the	traditional	i.t.

Figure 3. An example of alignment and merging.

3. IMPLEMENTATION

The new caption-editing tool is a Web application with a HTML5 and JavaScript-based front end that accesses Java Servlets through AJAX communications. The front end also uses JQuery, Bootstrap, and Adobe Flash applications written in ActionScript for the control functions of the video display. The speech recognition results were obtained from the IBM Attila speech recognition engine described in [13].

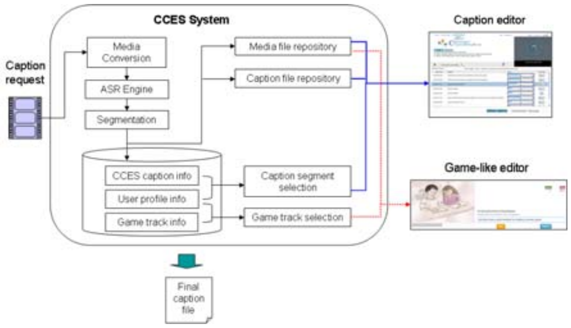
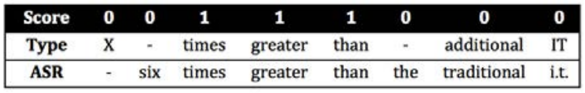


Figure 4. CCES and Editing Tools Architecture.

As mentioned in Section 2, the user interface of the original CCES editing tool was focused on creating an intuitive and easy-to-use caption editor with expert users for final quality verification. The new game-like system is intended to engage and sustain the interest of multiple non-expert users in caption creation. Both interfaces share video files, ASR results, segmentation details, and updated caption text information, per the system architecture shown in Fig. 4. Once the content owner uploads the media into the system, it is converted to the appropriate audio codec and sent to the speech recognition engine. The CCES receives the ASR results from the speech recognition engine and converts them into the caption format synchronized with the video. Then the system segments the generated caption file and corresponding video into chunks, and stores their information in a database. The owner of the content is given the option to choose which of the two methods is used for obtaining captions from the crowd.

Score Calculation: In Type mode, each user's submitted transcription is aligned with the ASR result. In Fix mode, this would discourage participants from making any changes to the proposed text since that would negatively affect their score. Therefore, in Fix mode, the comparison is performed with another user's transcription, created in Type mode, instead of with the ASR text. After the alignment, exact string comparison is used for each word pair to calculate the score, as illustrated in Fig. 5. Misspelled words and capitalization differences are considered as unmatched and the corresponding fluctuation in score provides feedback to the users.



Score	0	0	1	1	1	0	0	0
Type	X	-	times	greater	than	-	additional	IT
ASR	-	six	times	greater	than	the	traditional	i.t.

Figure 5. An example of scoring in Type mode.

Time-Limit Calculation: Sobhi et al. [12] found that it takes approximately 8 times the length of a video on average for non-experts to add captions to it. We increased this number to 9 as a starting point for our initial implementation and used it as the multiplier to calculate the time limit for editing the transcription of each video segment in the Type and Fix modes. Further investigation of the optimal multiplier is needed.

4. PILOT EXPERIMENT

We pre-populated our test system with video content from the Media Library of the authors' company, where base truth captions were already available from experts. This allowed us to calculate the accuracy of the collected results against the correct captions. A total of 578 English video segments, from 2 to 10 seconds long, were chosen from multiple videos spanning diverse categories such as narrations, tutorials, and speeches. The videos included multiple speakers in various recording environments.

The proposed editing tool was advertised to a small group of 100 people seeking voluntary participation. Out of these, 42 submitted captions for at least one video segment either in Type or Fix mode. The evaluation in this initial phase focused on the accuracy of the collected transcriptions.

4.1 Accuracy of Preliminary Collected Data

In this preliminary experiment each of the 578 video segments was considered with only two suggestions for its caption. The first one was obtained in Type mode and the second was in Fix mode. A limitation of this preliminary study is the small number of suggestions per video segment. We believe that obtaining more accurate transcriptions requires input from more than 2 participants per segment. However, these results can be viewed as a baseline for our editing tool. To evaluate the accuracy of the submitted transcriptions and the merged outcome, we compared the Word Error Rate (WER) to the ASR results given by the speech recognition engine. The WER metric provides strong agreement with the deaf and hard of hearing participants in evaluating the accuracy of the transcriptions [11] and is measured as the number of insertions (I), deletions (D), and substitutions (S) over the total length of the accurate transcript (N) on a word basis:

$$WER = \frac{I + D + S}{N}$$

In our calculations, all of the transcriptions were converted to lower case and the punctuation was ignored. Since the number of words for each short video clip varied from 2 to 24 words, the WER improvement is calculated for the aggregated content from all of the segments (a total of 6,501 words). We found that the overall WER dropped from 20.7% in ASR to 16.0% in the merged results. Figure 6 shows the distribution of the WER and the errors in the captions of the video segments as boxplots with whiskers at the 1.5 IQR (inter-quartile range). To aid the comparison, median values are added as labels at the top of each plot and the mean values are identified with a star at (22%, 33%, and 16%) for the WER and at (2.35, 3.8, and 1.82) for the number of errors. We observe that the merged results have a smaller variance in both cases. This suggests that the proposed editing tool should work better if input from more than 2 users is considered. While the transcriptions that were submitted in Type mode seems to be worse than the ASR results, as shown in Fig. 6, they can prevent errors from the ASR that may mislead users in the Fix mode. If the ASR result is almost perfect, users tends to trust that the suggestion is correct, thus missing necessary corrections, e.g. of the word "blocks" to "blogs". In Type mode, where no suggestions are provided, users might detect the word "blogs" correctly, as shown in a few cases by our data.

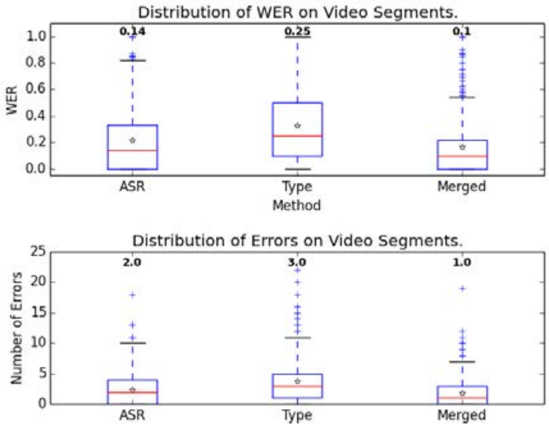


Figure 6. WER and Error Distribution for Video Segments.

We observed that 20% of the participants submitted captions for at least 12 video segments. Overall, the participants skipped the 'type' mode 40 times and the 'fix' mode 39 times in a total of 71 distinct video segments. In 'type' mode the participants submitted their transcriptions while 0-35 (mean 6.32) seconds remained in the countdown clock, and in 'fix' mode 0-48 (mean 19.76) seconds remained.

5. WORK IN PROGRESS

The caption-editing tool described in this article introduces game-like features into a crowdsourced video captioning system with promising preliminary results based on combining input from speech recognition and a crowd of non-experts in listen-and-type and listen-and-fix modes. Both the interface and the underlying mechanisms for obtaining accurate captions can be significantly improved with additional game mechanisms and game design techniques that fall under the general motif of Games With a Purpose (GWAP) [14]. Here are some of the more promising ideas for future research:

Gamification: CCES is a collaborative captioning system whose interface is intuitive and easily used by non-experts. When a request for captioning of video content is sent to the crowd, the system could be enhanced with incentives to motivate users to participate in the task. We are investigating if a GWAP approach could be adopted given that captioning is intrinsically limited in entertainment value. For example, a two-player game may attract people to continue playing a game that generates video captions in the background without other economic incentives.

Ensuring Accuracy: The quality of the results is always an important consideration for crowdsourced tasks, since participants are usually recruited from the public. CCES and its initial caption-editing tool used trained experts to assure high quality. The editors contributed the initial transcriptions, which were reviewed and corrected by the experienced users. Even though the reviewing process can also be segmented and distributed to a crowd, it still needs trained humans to improve the quality. We are investigating task completion criteria in a system that combines input from multiple participants and automatically determines if the task results are accurate with a given probability.

6. CONCLUSION

We are working on an online video captioning editor that incorporates game-like elements to motivate a crowd of non-experts, ultimately leading to accessibility benefits for people who are deaf or hard-of-hearing. The presented application obtains transcriptions from multiple users on video segments that are a few seconds long. Users submit their transcriptions against a countdown timer either by listening to and typing the content of the audio or by editing the ASR result, and thus earn points. Transcriptions from multiple players on the same video segment are then aligned and merged.

Preliminary results on a small crowd of 42 participants with 578 short video segments show that the accuracy of the merged captions, as measured by Word Error Rate, improved from 20.7% with the ASR to 16.0%, limited to calculations on inputs from 2 users per video segment. We are currently investigating the relationship of the number of submitted transcripts per segment and the accuracy of the merged results, as well as alternative merging methods and task completion criteria.

Crowd engagement, one of the major considerations in crowdsourcing, becomes more challenging when the task has inherently limited entertainment value, as with video captioning, where a user's input is closely bound to the media and there is little room for creativity. We are currently considering a two-player game as an improvement to the proposed system and other non-monetary incentive mechanisms to address the limitations.

7. ACKNOWLEDGMENTS

This research was partially supported by the Japan Society and Technology Agency (JST) under the Strategic Promotion of Innovative Research and Development program. Authors gratefully acknowledge the helpful feedback from the members of the IBM Tokyo Accessibility Research Group and Human Ability and Accessibility Center.

8. REFERENCES

[1] Amazon Mechanical Turk. http://www.mturk.com.

[2] Automatic captions for YouTube videos. http://googleblog.blogspot.com.

[3] CastingWords. https://castingwords.com.

[4] Gruenstein, A., McGraw, I., and Sutherland, A. 2009. A self-transcribing speech corpus: collecting continuous speech with an online educational game. In *SLaTE Workshop*.

[5] Goto, M., and Ogata, J. 2011. PodCastle: Recent Advances of a Spoken Document Retrieval Service Improved by Anonymous User Contributions. In *INTERSPEECH 2011*, 3073-3076.

[6] Kirkland, E., Byrom, E., MacDougall, M., and Corcoran, M. 1995. The Effectiveness of Television Captioning on Comprehension and Preference. *American Educational Research Association*, 1995 Annual Meeting, San Francisco, CA.

[7] Krajka, J. 2013. Audiovisual Translation in LSP–A Case for Using Captioning in Teaching Languages for Specific Purposes. *Scripta Manent* 8(1), 2-14.

[8] Lasecki, W. S., Miller, C., Sadilek, A., Abumoussa, A., Borrello, D., Kushalnagar, R., and Bigham, J. P. 2012. Real-time captioning by groups of non-experts. In *Proc. UIST 2012,* ACM, 23-34.

[9] Liem, B., Zhang, H., and Chen, Y. 2011. An Iterative Dual Pathway Structure for Speech-to-Text Transcription. In *Human Computation*.

[10] Meyer, M. J., and Lee, Y. B. B. 1995. Closed-Captioned Prompt Rates: Their Influence on Reading Outcomes. Office of Special Education and Rehabilitative Services.

[11] Naim, I., Gildea, D., Lasecki, W. S., and Bigham, J. P. 2013. Text Alignment for Real-Time Crowd Captioning. In *Proc. NAACL-HLT 2013*, 201-210.

[12] Sobhi, A., Nagatsuma, R., and Saitoh, T. 2012. Collaborative Caption Editing System–Enhancing the Quality of a Captioning and Editing System. In *Proc. of the 28th Annual International Technology and Persons with Disabilities Conference (CSUN)*.

[13] Soltau, H., Saon, G., and Kingsbury, B. 2010. The IBM Attila speech recognition toolkit. In Spoken Language Technology Workshop (SLT), IEEE, 97-102.

[14] von Ahn, L., and Dabbish, L. 2008. Designing games with a purpose. *Comm. ACM*, 51(8), 58-67.

[15] Wald, M. 2013. Concurrent Collaborative Captioning. In *Proc. SERP 2013*

[16] Winke, P., Gass, S., and Sydorenko, T. 2010. The effects of captioning videos used for foreign language listening activities. *Language Learning & Technology*, 14(1), 65-8.