Legion Scribe: Real-Time Captioning by Non-Experts

Walter S. Lasecki[1], Raja Kushalnagar[2], and Jeffrey P. Bigham[3]

Computer Science, ROC HCI[1]
University of Rochester
wslasecki@cs.rochester.edu

Computer Science, NTID[2]
Rochester Institute of Technology
rskics@rit.edu

HCII[3]
Carnegie Mellon University
jbigham@cmu.edu

ABSTRACT

The promise of affordable, automatic approaches to real-time captioning imagines a future in which deaf and hard of hearing (DHH) users have immediate access to speech in the world around them my simply picking up their phone or other mobile device. While the challenges of processing highly variable natural language has prevented automated approaches from completing this task reliably enough for use in settings such as classrooms or workplaces [4], recent work in crowd-powered approaches have allowed groups of non-expert captionists to provide a similarly-flexible source of captions for DHH users. This is in contrast to current human-powered approaches, which use highly-trained professional captionists who can type up to 250 words per minute (WPM), but also can cost over $100/hr. In this paper, we describe a real-time demo of Legion:Scribe (or just "Scribe"), a crowd-powered captioning system that allows untrained participants and volunteers to provide reliable captions with less than 5 seconds of latency by computationally merging their input into a single collective answer that is more accurate and more complete than any one worker could have generated alone.

Categories and Subject Descriptors

H.5.2 [Information interfaces and presentation]: User Interfaces. - Graphical user interfaces.

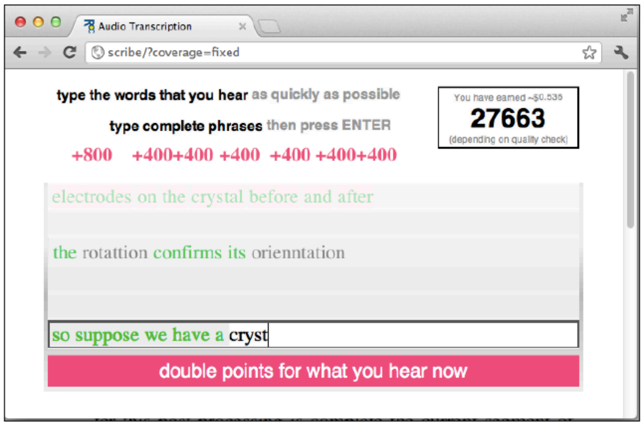
General Terms

Design, Human Factors

Keywords

Captioning, speech-to-text, real-time human computation, deaf, hard of hearing, crowdsourcing

ASSETS 2014 – Captioning Competition, October 20–22, 2014, Rochester, NY.



type the words that you hear as quickly as possible

type complete phrases then press ENTER

electrodes on the crystal before and after

the rotattion confirms its orienntation

so suppose we have a cryst

Figure 1: The worker interface encourages workers to type audio by locking in words soon after they are typed. To encourage typing specific segments, visual and audio cues are given, and the volume of the audio is reduced during off periods, while rewards are increased for on periods.

1. LEGION SCRIBE

Real-time captioning provides word-for-word transcription of spoken content with a latency of less than 5 seconds. This can provide as an access to live events for deaf and hard of hearing people, and help hearing users keep track of their place and avoid missing content they might have otherwise missed.

Currently, the only reliable solution in real-life situations with differing speaker and environments is to hire a professional stenographer who provides captions for over $100/hour or more, depending on their skill level.

Real-time captioning is difficult because even fast typists cannot type at the 250 words per minute (WPM) necessary to keep up with natural speech (stenographers train to use special keyboards and input phonemes instead of individual letters, resulting in much lower typing rate requirements). Automatic speech recognition (ASR) only captures about 40% of the speech in real settings and confuses readers by making seemingly random errors [4].

Legion:Scribe [2] allows a small group (often as low as 3-5 people) of people who can hear and type at a ordinary rate to collectively caption speech in real-time, for 20-30% the cost of an expert stenographer. Because of the relative frequency of people who know how to type at a nominal rate compared with those who are trained stenographers, Scribe presents a captioning service that can be available at a moment's notice, rather than at least 24 to 48 hours, as is the case with professionals.

Scribe can now be deployed using a server architecture that allows workers to simply connect to a web page, enter an automatically generated easy-to-remember pass phrase (e.g., "apple", or "hard hat"), and immediately be routed to an audio stream with appropriate indicators of when to type what they hear. For end-users, they simply arrive at a viewing page, enter a pass phrase, and see a display of the captions as they are entered.

By using multiple captionists, it is less likely that the system falls very far behind the speaker. This means that compared to a single stenographer, the flow of the text is more consistent. This likely helps users read content more easily, and is one reason why the captions produced by Scribe, while not perfect, can actually be preferred to professionals [4]. The interface also includes pausing and highlighting controls that allow individual readers to control how they access the captions themselves, which has been shown to help readers better follow content [1].

2. ENHANCING INDIVIDUAL ABILITIES

Scribe asks multiple people to type what they hear into the interface shown in Figure 1 and then merges the partial captions back together using a multiple sequence alignment algorithm. How many workers are required to cover everything a speaker says depends on the speaker's rate and the typing rate of the captionist. However, typically between 3 and 5 people can accomplish the task effectively.

In our experiments, we also recruited workers cheaply and on-demand from crowdsourcing marketplaces like Amazon Mechanical Turk, Mobile Works, and oDesk. These non-expert captionists can be recruited on-demand for as long or as short as the user needs, allowing the workforce used for powering these service even more flexible and available than a set of rare, highly-skilled professionals could be.

Our approach leverages both the fact that people are available on-demand, often as volunteers, as well as that by properly coordinating group work, group systems can outperform even the highest-skilled individual in the group. We have also shown that by systematically slowing down and speeding up the audio for individual workers we can improve both precision and recall by more than 10% [3]. This is the *TimeWarp* approach to real-time human computation. Our interface (Figure 1) coordinates different workers so they type different portions of the streaming audio while maintaining the context of all of the speech. Scribe currently comes close to the performance of stenographers in terms of *coverage,* how many of the words in the ground truth appear in the final output stream, and *precision*, how many of the output words are correct. Tests have shown Scribe is currently able to reach a precision of 84.8% of that of a professional. We expect that over time Scribe will become even more competitive and might be able to even surpass the performance of stenographers in terms of both coverage and precision. Additionally, experiments with TimeWarp also indicate the potential for improvements in latency, meaning multiple workers might also be able to outperform a single expert even in terms of speed, given the right workflow.

3. DEMO

Our demo will have three components. First, our remote captionists (volunteers and students with no previous captioning training) will listen to audio and provide captionists via Scribe. Second, readers will be able to access the real-time captions generated by Scribe via both an on-screen projection, and a viewing page with playback controls that each reader can use independently. Finally, we will allow the volunteers from the audience to participate in generating captions shown in a second public display.

4. SUMMARY

We have presented **S**cribe, a reliable human-powered approach for providing on-demand real-time captioning at low cost. Scribe uses multiple individual captionists, each of whom type a piece of what they hear, and then stitches the partial captions back together automatically. Scribe performs competitively with current approaches, but for a fraction of the cost to operate. Our demo will allow members of the audience to both view and, if they choose, help collectively produce captions.

5. REFERENCES

[1] W. S. Lasecki, R. Kushalnagar, J. P. Bigham. Helping students keep up with real-time captions by pausing and highlighting. In Proceedings of W4A 2014. Article 39, p1-8. DOI=10.1145/2596695.2596701

[2] W. S. Lasecki, C. D. Miller, A. Sadilek, A. Abumoussa, D. Borrello, R. Kushalnagar, and J. P. Bigham. Real-time captioning by groups of non-experts. In *Proceedings of UIST 2012*. p23-33. DOI=10.1145/2380116.2380122

[3] W. S. Lasecki, C. D. Miller and J. P. Bigham. Warping Time for More Effective Real-Time Crowdsourcing. In *Proceedings of CHI 2013*. p2033-2036. DOI= 10.1145/2470654.2466269

[4] R.S. Kushalnagar, W.S. Lasecki, J.P. Bigham. Accessibility Evaluation of Classroom Captions. ACM Transactions of Accessible Computing (TACCESS). 5,3, Article 7. DOI=10.1145/2543578