CARTGPT: Improving CART Captioning using Large Language Models

LIANG-YUAN WU

University of Michigan, Ann Arbor, lyuanwu@umich.edu

ANDREA KLEIVER

University of Michigan, Ann Arbor, akleiver@umich.edu

DHRUV JAIN

University of Michigan, Ann Arbor, profdj@umich.edu

Communication Access Realtime Translation (CART) is a commonly used real-time captioning technology used by deaf and hard of hearing (DHH) people, due to its accuracy, reliability, and ability to provide a holistic view of the conversational environment (*e.g.*, by displaying speaker names). However, in many real-world situations (*e.g.*, noisy environments, long meetings), the CART captioning accuracy can considerably decline, thereby affecting the comprehension of DHH people. In this work-in-progress paper, we introduce *CARTGPT*, a system to assist CART captioners in improving their transcription accuracy. CARTGPT takes in errored CART captions and inaccurate automatic speech recognition (ASR) captions as input and uses a large language model to generate corrected captions in real-time. We quantified performance on a noisy speech dataset, showing that our system outperforms both CART (+5.6% accuracy) and a state-of-the-art ASR model (+17.3%). A preliminary evaluation with three DHH users further demonstrates the promise of our approach.

CCS CONCEPTS • Human-centered computing~Accessibility~Empirical studies in accessibility • Human-centered computing~Accessibility~Accessibility technologies

Additional Keywords and Phrases: Accessibility, Deaf and hard of hearing, real-time captioning.

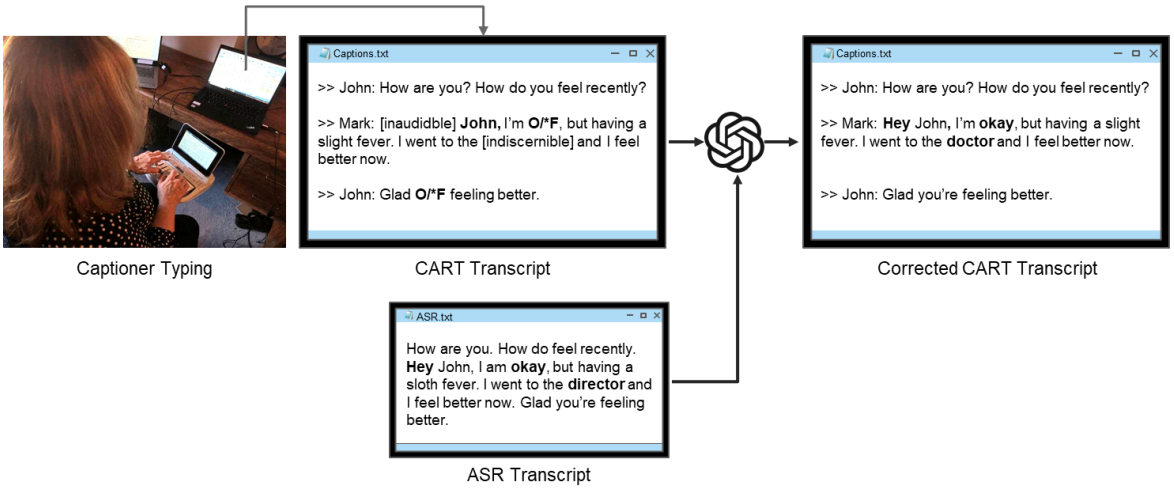

Figure 1: Our CARTGPT system takes in the CART transcript and the ASR transcript as input and produces the corrected CART transcript. The corrected words are highlighted in **bold**.

1 INTRODUCTION

Communication Access Real-time Translation (CART) or real-time captioning, employs a shorthand keyboard to transcribe spoken content (Figure 1) [9,15,23]. CART captioning is a preferred captioning technology by DHH people, not only due

to its ability to produce highly accurate captions that considerably outperform ASR [24,25], but also due to its ability to provide a holistic view of the conversation by displaying speaker names, speech tone, and important contextual cues in captions (*e.g.,* someone laughing or doorbell ringing) [9,24]. However, in many real-world situations—such as fast speakers, long meetings, and noisy environments—the quality of the transcription can significantly decline [1,14], thereby affecting the comprehension of deaf and hard of hearing (DHH) users. In this work, we examine computational approaches to improve CART generated captions.

While prior work has not examined improving CART captions, researchers have long examined methods to improve the accuracy of automatic speech recognition (ASR) by training larger AI models [4,11,21], collecting more varied datasets [7,16,17], and by including additional specialized algorithms (*e.g.,* noise reduction algorithms [3,13], or large language models (LLMs) [12]) in the ASR pipeline. Some HCI work has also explored real-time editing of ASR text by untrained humans [6] or foregoing ASR altogether by using crowdsourcing approaches [10]. However, these approaches still fail to perform at par with trained CART captioners, especially in real-world situations [24,25]. We sought to investigate if we can even further improve the accuracy of CART captioning to enhance the conversational experience of DHH users who rely on captions.

2 FORMATIVE STUDY

We began with a formative interview with 10 CART captioners to understand how and what kind of errors may appear in captions and to gauge their interest in using future technology to improve their captions. We recruited certified CART captioners (7 women, 2 men, 1 non-binary) through captioning agencies (*e.g.,* CaptionFirst, QuickCaption), email lists, and snowball sampling. Participants were on average 37.4 years old (*SD*=13.6, *range*=25-57) and had several years of experience with captioning: two had >20 years, three had 15-20 years, two had 10-15 years, and three had 2-10 years. Interview sessions were conducted online over Zoom and lasted approximately one hour. We compensated the participants with USD 25. The interview recordings were transcribed and analyzed using applied analysis thematic approach, which contained iterative coding by one researcher, independent coding using the final codebook by another researcher, inter-coder agreement calculation (Krippendorff's alpha [8] was 0.85), and disagreement resolution.

We found that unideal conditions did affect captioning, with participants mentioning various factors for reduced quality such as extremely technical conversation topics (*e.g., "a lecture on Algorithms"* - P8) (*N*=8), long meetings (*N*=8), noisy environments (*N*=8), unclear, rapid, or accented speech (*N*=7), and unideal seating arrangements (*e.g.,* captioner seated far from the speaker) (*N*=6). To investigate how future NLP technology can provide support, we asked the captioners what types of errors appear in the captions. We categorized the errors into four types based on how they appear in a transcript.

The first category includes word or phrase omissions arising from unclear or accented speech. Captioners signify these errors using two standard keywords in their transcript, "[inaudible]" and "[indiscernible]", as placeholders for the missing words or phrases. The keyword "[inaudible]" is used when speech is completely inaudible due to low volume, microphone issues, or the speaker's distance from the captioner. In contrast, "[indiscernible]" denotes speech that is audible, but is indistinguishable, for example, due to accents.

The second category covers word or phrase omissions originating from factors other than inaudible or indiscernible speech such as noise, technical content, or rapid speakers. These errors are indicated by a special character "(?)".

The third and fourth error categories, untranslate and mistranslate, result from typing mistakes from captioners. Untranslate errors occur when a wrong key combination is pressed (called a 'mistroke'). For example, keys 'SPBRO/E' translate to the prefix "intro-" but if key A is pressed instead of E, the transcription shows raw 'SPBRO/A". These errors are distinctly identifiable in the transcript owing to their appearance with adjoining capital letters and special characters.

In contrast, mistranslate errors occur when a mistroke results in an actual word or phrase present in the dictionary, but it differs from the intended one. For example, keys "LAB/DOER" translate to "Labrador", but "LAB/DOERZ" was pressed incorrectly, which resulted in "Lab board of directors". Mistranslates are not readily distinguishable in a transcript.

When asked about whether they would be interested in using computational technology to improve their captioning, all captioners were supportive of the idea, and recommended use of ASR (N=9), or even LLMs (*e.g.*, *"ChatGPT"*) (N=3) to supplement their captioning. For example, P4 said: *"I know I can type much better than automatic [speech recognition], but sometimes like in long meetings or if someone is speaking like really fast, I can't keep up, and I would much like that AI can help fill missing words."*

3 THE CARTGPT SYSTEM

Informed by the above findings, we investigated the potential of automated approaches to improve CART captions. Our aim was not to replace CART captioning but work in tandem with the captioners to achieve even higher accuracy than CART alone. Therefore, we built a system called *CARTGPT*, which uses errored CART captions along with the ASR transcript of the conversation to generate corrected CART captions in real-time. CARTGPT employs a large language model that searches for specific errors in the captions (three out of four error types reported in our formative study above) and uses the corresponding ASR script along with the context of the conversation to predict replacement words. The replacement words are not merely copied from the ASR transcript but are also corrected if needed, since the ASR transcript will likely have errors too. Figure 1 demonstrates an example use case, where the word "director" from the ASR output was corrected to "doctor" based on the learned conversational context. We explain the error correction process below.

3.1 Searching for Error Keywords

First, CARTGPT streams the CART text in real time, searching for specific keywords to be replaced. Based on the formative study findings, we search for the following three readily identifiable CART errors:

1. Omissions due to inaudible or unclear speech, which are specified using keywords "[inaudible]" or "[indiscernible]"
2. Other omitted words or phrases, which are specified using "(?)"
3. Untranslate errors, which appear in CART text with adjoining capital letters and special characters (*e.g.*, "we were able to **O/*F** the process…" or "After an **SPBRO/A** of the lesson,")

As the first initiative in this area, we excluded the fourth error type, mistranslate errors, because they may not be readily identifiable in captioned text. For example, the phrase *"The labrador went for a walk"* mistranslated to *"The lab board of directors went for a walk"* could still make sense in context. We leave this for future work to investigate further. Before the next step, all errors are converted to the placeholder character '[…]' and handled uniformly.

3.2 Using the LLM to Generate Replacement Text

Once the prototype detects an error keyword, it calls an LLM, *GPT-4* [26], to generate plausible words to replace the error keywords. The model uses the corresponding ASR transcript (obtained from the OpenAI Whisper [27] model) along with context of the spoken content to find the replacement words. To enable the model to learn the conversational context, we provide it with two preceding paragraphs of CART text. Specifically, we use the following prompt:

> *You are correcting a CART transcript. Please replace the text "[…]" with the words or phrases that best fit the context. Do not change anything else. Use the following preceding text and the ASR transcript of the same conversation to learn from the context:*

Preceding text: [Two preceding paragraphs of CART transcript including the current paragraph]
ASR transcript: [ASR text]

To arrive at this prompt, we conducted heuristic experiments with two speech benchmarks: *LibriSpeech* [1] and *TED-LIUM* [2]. Other prompting strategies, such as providing no context (*i.e.*, zero-shot) or providing much more context (*e.g.*, five paragraphs), performed marginally poorly than our current prompting strategy; echoing past work [12].

3.3 Replacing the Error Keywords with Generated Text

Our internal experiments revealed that although the LLM replaces all the error keywords as intended, it also occasionally substitutes other words or phrases in an attempt to simplify the text–a behavior that reflects prior findings [12,19]. Thus, we run a post-processing step, where we revert any other replaced words in the CART captions to their original text.

4 EXPERIMENTAL EVALUATION

Before evaluating with our target users, it was necessary to demonstrate that our approach indeed improves CART captioning. Therefore, we quantified performance on a noisy real-world speech dataset we collected. We collected speech files from four publicly available benchmarks: TED-LIUM [7], Patient-Physician medical interviews [2], MIT OCW [28], and CallHome [29]. Collectively, these benchmarks contain speech spanning multiple domains (*e.g.*, medical, computer science, common conversation topics) and conversation styles (*e.g.*, lectures, group meetings, one-on-one meetings) along with their ground truth transcripts. From each benchmark, we randomly selected files to span about 10 hours of content (*e.g.*, from the TED-LIUM dataset which contains recordings of approx. 15 mins, we chose 40 files). In total, our dataset spans 39.7 hours.

CART captioning is usually very accurate (*e.g.*, upwards of 98% [1,15]), but the accuracy declines in unideal conditions (*e.g.*, long meetings, noisy environments). Thus, to emulate unideal real-world conditions, we added noise to our dataset by mixing each audio file with one of the six environmental noises we collected at varying signal-to-noise ratios: HVAC, babbling, urban ambience, medical equipment running, exhibition hall background, and lecture hall acoustics.

To generate CART transcripts, we hired three CART captioners from [anonymized]. To mitigate individual differences in captioning, we asked each captioner to transcribe every audio file. To obtain the ASR transcript, we used OpenAI's Whisper model. We then used our CARTGPT system to generate the corrected transcript for each captioner's transcripts. For measurement, we used the Word Error Rate (WER) metric to compare the final transcripts to their ground truths. Any non-verbal contextual cues inserted by the captioners (*e.g.*, speaker names) were excluded from the calculation.

We found that average accuracy of CARTGPT was 89.0% (*i.e.*, *WER*=0.110) (*SD*=5.8%), an improvement of 5.6% over CART (83.4%, *SD*=7.9%) and 17.3% over ASR (71.7%, *SD*=12.9%). This improvement was significant; a pairwise t-test across all transcripts yielded t_{111}=8.8, p<.001 for CARTGPT *vs.* CART and t_{111}=12.9, p<.001 for CARTGPT *vs.* ASR.

Across different conversational topics (*e.g.*, medical, computer science, food, weather), we found that the improvement was more pronounced for technical topics (*e.g.*, medical or computer science, increase of +5.8% over CART) compared to casual topics (*e.g.*, weather, food, increase of +3.2% over CART), likely because technical topics pose a greater challenge for the captioners to comprehend.

5 PRELIMINARY USER FEEDBACK

We are undergoing studies with DHH people, and thus far, have completed three participants (two women, one man). We recruited our participants through social media and our email lists. They were on average 34.3 years old and identified as Deaf, hard of hearing, and deaf. Two participants had severe hearing loss, and one had profound hearing loss. The IRB-approved study was conducted in our research lab and lasted 50 mins. We compensated the participants with USD 50.

During the study, we recruited a CART captioner who generated the captions on a computer running Open AI Whisper model in a separate application. Captions from both the captioners and the Whisper model were exported in real-time to separate text files. These files were then processed by our CARTGPT system to generate the corrected captions which were shown to the user. Also, the captions from the captioner were shown in a separate window placed side-by-side with the CARTGPT captions.

We asked the captioner to transcribe two pre-recorded conversations played on a speaker: a fast-faced 15-min technical computer science lecture and a 15-min casual conversation on weather. After each conversation, participants were asked to rate their comprehension on a scale of 1 to 5 and provide rationale for their rating. At the end of the study, we asked participants for overall thoughts on our approach. Sessions were audio recorded and later transcribed for analysis. For the Deaf participant, we also recruited a sign language interpreter. For the analysis, two researchers worked together to summarize the data and reveal initial themes.

Our initial findings show that all three participants prefer our approach (average comprehension=4.3/5) over the traditional CART captions (average comprehension=3.7/5). When asked for subjective preferences, participants claimed that the replaced words *"made sense"* (P2), and *"increased my understanding of whatever was being said"* (P1). They also reported that they did not observe a visible time difference between the two texts, confirming that our approach is able to operate in real-time.

6 DISCUSSION AND CONCLUSION

We detailed the design and evaluation of CARTGPT, a system to address CART captioning errors using the accompanying ASR transcript and an LLM model. Taken together, our evaluations provide initial evidence that our prototype *can* improve CART captioning in diverse unideal acoustic environments in real-time without domain-specific training. We are currently undergoing further trials with our users, and plan to recruit a total of 10-12 DHH participants. Besides this study, we see several other opportunities for future work:

Handling Other Error Types. Our prototype operates on the text generated by CART and ASR and does not interact with the speaker audio. Therefore, it only addresses the error types that are identifiable in the CART transcripts (omissions due to intelligible speech, other omissions, and untranslates). Future work should investigate whether supplementing our LLM prototype with audio embeddings extracted from speech could further improve the accuracy of captions, allowing it to identify and correct the fourth error type (mistranslate) and beyond.

Human-in-the-Loop. Another interesting area to investigate is human-in-the-loop based approaches [18] where DHH end-users or the CART captioners can provide valuable feedback to strengthen the model. For example, these users can supply additional contextual cues (*e.g.,* the location of use) and actively participate in the correction process (*e.g.,* by accepting or rejecting an LLM suggested modification) to further improve the performance of our technique.

Domain-Specific Models. Our work used a general purpose LLM to handle a variety of conversational topics and style showing the adaptability and versatility of our prototype in diverse contexts. However, specialized trained LLMs (*e.g.,* for education or healthcare) can improve text accuracy and coherence even further [5,22], and should be examined for implementation in specific settings (*e.g.,* classrooms or hospitals).

Privacy and On-Device Implementation. Our prototype interfaces with GPT-4 running on the cloud, which could raise privacy concerns, particularly in sensitive environments (*e.g.,* healthcare). Recently, low-resource models such as Alpaca [20] have emerged, designed to fit on edge devices. As these compact models evolve, they should be examined for on-device deployment of our prototype.

REFERENCES

[1] Gregory J Downey. 2008. Closed captioning: Subtitling, stenography, and the digital convergence of text with television. JHU Press.

[2] Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, Thomas Lo, and Christopher W. Smith. 2022. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. Scientific Data 9, 1: 313. https://doi.org/10.1038/s41597-022-01423-1

[3] Kanika Garg and Goonjan Jain. 2016. A comparative study of noise reduction techniques for automatic speech recognition systems. In 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2098–2103.

[4] Matteo Gerosa, Diego Giuliani, Shrikanth Narayanan, and Alexandros Potamianos. 2009. A review of ASR technologies for children's speech. In Proceedings of the 2nd Workshop on Child, Computer and Interaction, 1–8.

[5] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) 3, 1: 1–23.

[6] Rebecca Perkins Harrington and Gregg C Vanderheiden. 2013. Crowd caption correction (ccc). In Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility, 45.

[7] François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. 2018. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18--22, 2018, Proceedings 20, 198–208.

[8] Klaus Krippendorff. 2018. Content analysis: An introduction to its methodology. Sage publications.

[9] Judy Larson and others. 1999. CART (Communication Access Realtime Translation). PEPNet Tipsheet. PEPNet-Northeast.

[10] Walter S Lasecki, Christopher D Miller, Raja S Kushalnagar, and Jeffrey P Bigham. 2013. Legion Scribe: Real-Time Captioning by the Non-Experts. In 10th International Cross-Disclipinary Conference on Web Accessibility (W4A).

[11] Bo Li, Anmol Gulati, Jiahui Yu, Tara N Sainath, Chung-Cheng Chiu, Arun Narayanan, Shuo-Yiin Chang, Ruoming Pang, Yanzhang He, James Qin, and others. 2021. A better and faster end-to-end model for streaming asr. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5634–5638.

[12] Rao Ma, Mengjie Qian, Potsawee Manakul, Mark Gales, and Kate Knill. 2023. Can Generative Large Language Models Perform ASR Error Correction? arXiv preprint arXiv:2307.04172.

[13] Andrew Maas, Quoc V Le, Tyler M O'neil, Oriol Vinyals, Patrick Nguyen, and Andrew Y Ng. 2012. Recurrent neural networks for noise reduction in robust ASR. INTERSPEECH.

[14] Somang Nam, Maria Karam, Christie Christelis, Hemanshu Bhargav, and Deborah I Fels. 2023. Assessing subjective workload for live captioners. Applied Ergonomics 113: 104094.

[15] National Association of the Deaf (NAD). Communication Access Realtime Translation. Retrieved April 7, 2018 from https://www.nad.org/resources/technology/captioning-for-access/communication-access-realtime-translation/

[16] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5206–5210.

[17] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. arXiv preprint arXiv:2012.03411.

[18] Gonzalo Ramos, Christopher Meek, Patrice Simard, Jina Suh, and Soroush Ghorashi. 2020. Interactive machine teaching: a human-centered approach to building machine-learned models. Human--Computer Interaction 35, 5–6: 413–451.

[19] Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, 1–7.

[20] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. Retrieved from https://crfm.stanford.edu/2023/03/13/alpaca.html

[21] Dong Wang, Xiaodong Wang, and Shaohe Lv. 2019. An overview of end-to-end automatic speech recognition. Symmetry 11, 8: 1018.

[22] Yue Zhang, Leyang Cui, Deng Cai, Xinting Huang, Tao Fang, and Wei Bi. 2023. Multi-Task Instruction Tuning of LLaMa for Specific Scenarios: A Preliminary Study on Writing Assistance. arXiv preprint arXiv:2305.13225.

[23] What is real-time captioning? | UW DO-IT. Retrieved August 12, 2022 from https://www.washington.edu/doit/what-real-time-captioning#:~:text=Captions%2C composed of text%2C are,as an event takes place.

[24] Captions: Humans vs Artificial Intelligence: Who Wins? | Equal Entry. Retrieved September 14, 2023 from https://equalentry.com/caption-videos-human-vs-automatic-captions/

[25] Live Professional Captions vs. CART Captioning. Retrieved September 14, 2023 from https://www.3playmedia.com/blog/live-professional-captions-vs-cart-captioning-whats-the-difference/

[26] GPT-4 | OpenAI. Retrieved July 2, 2024 from https://openai.com/index/gpt-4/

[27] Introducing Whisper | OpenAI. Retrieved July 2, 2024 from https://openai.com/index/whisper/

[28] MIT OpenCourseWare | Free Online Course Materials. Retrieved September 13, 2023 from https://ocw.mit.edu/

[29] CALLHOME American English Speech - Linguistic Data Consortium. Retrieved September 13, 2023 from https://catalog.ldc.upenn.edu/LDC97S42