CARTGPT: Improving CART Captioning using Large Language Models

Liang-Yuan "Leo" Wu and Dhruv "DJ" Jain
University of Michigan, Computer Science and Engineering

Introduction

- Communication Access Realtime Translation (CART) is a real-time captioning technology that is preferred by Deaf and Hard-of-hearing (DHH) individuals.
- It is more accurate than automatic approaches (e.g., ASR) and provides a holistic view of the conversation (e.g., by displaying speaker names, emotional cues).
- However, there are still factors that degrade the accuracy of CART such as faster speaker, long meetings, and noisy environments.
- In this poster, we present CARTGPT, a system that can further improve the accuracy of CART using a combination of ASR and LLM approach.

Formative Study

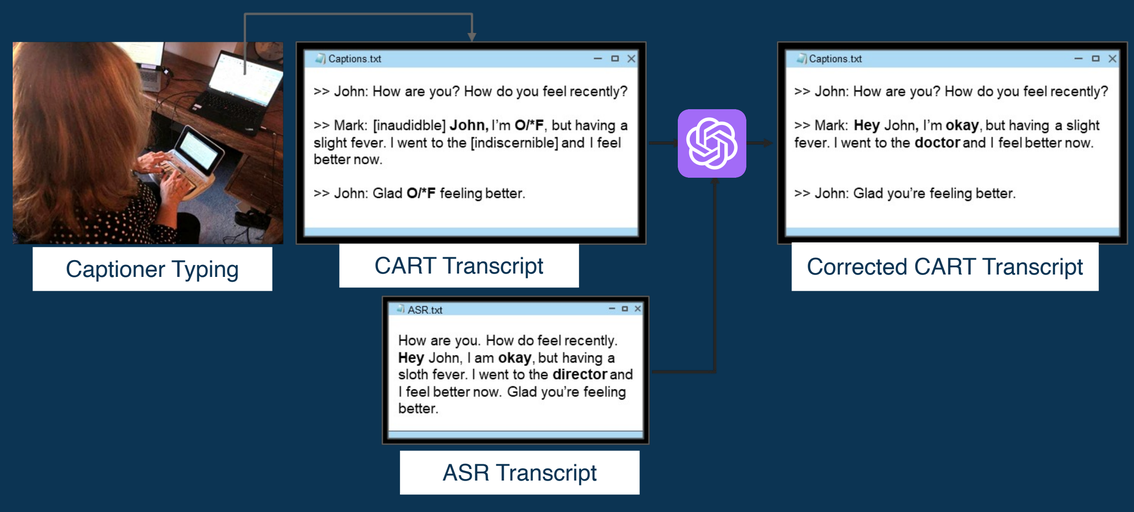
First, we interviewed with 10 professional CART captioners and concluded four categories of errors in CART captioning.

- Unclear or accented speech
- Noise, technical terms, rapid speaker
- Typing mistakes (mistroke)
- Mistranslate errors

To reduce these errors, all captioners were supportive of the idea to use automated approaches in tandem with their typing.

The CARTGPT System

1. Search for errors in CART
2. Use ASR along with an LLM (for alignment, contextual searching) to replace the errored words
3. Finally, post-process text to mitigate any hallucinations



You are correcting a CART transcript. Please replace the text "{cart_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: {paragraphs} ASR transcript: {asr_transcripts}

Evaluation & Results

We sampled approximately 10 hours of speech from four noisy speech datasets: TED-LIUM, Patient-Physician Conversation, MIT OCW and CallHome. We compared the word error rate (WER) of CART, ASR and our method, CARTGPT.

CARTGPT showed a significant 5.6% improvement over the original CART output and a significant 17.3% improvement over ASR.

CART	ASR	CARTGPT
83.4% (SD=7.9%)	71.7% (SD=12.9%)	**89.0%** (SD=5.8%)

We also performed a initial user study with 3 DHH participants. After experiencing both the traditional CART approach and our CARTGPT approach, participants commented that:

- CARTGPT improved their comprehension over traditional CART
- No visible processing delay was observed in CARTGPT vs. CART

Discussion and Future Work

We evaluated CARTGPT quantitively with a dataset and qualitatively with DHH people. We will continue our user study and plan to recruit a total of 12 participants. Several other directions for future work include:

- CARTGPT with Audio Embeddings: Supplement the LLM with audio information to correct mistranslate errors.
- Human-in-the-Loop: Iterative Feedback from users to strengthen the model.
- Domain-Specific Models: Specialized trained LLMs to handle specific conversation topics and contexts.
- Privacy and On-Device Implementation: Compact models that can run on-device to enhance privacy.

Leo's Website:
https://binomial14.github.io

CSE COMPUTER SCIENCE AND ENGINEERING UNIVERSITY OF MICHIGAN