CARTGPT: Improving CART Captioning using Large Language Models

Liang-Yuan “Leo” Wu and Dhruv 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., Automatic Speech Recognition (ASR)) and provides a holistic view of the conversation.
- However, there are still some factors that will degrade the accuracy of CART, including faster speaker, long meetings and noisy environments.
- In this poster, we present CARTGPT, a system that can further improve the accuracy of CART automatically.

Formative Study

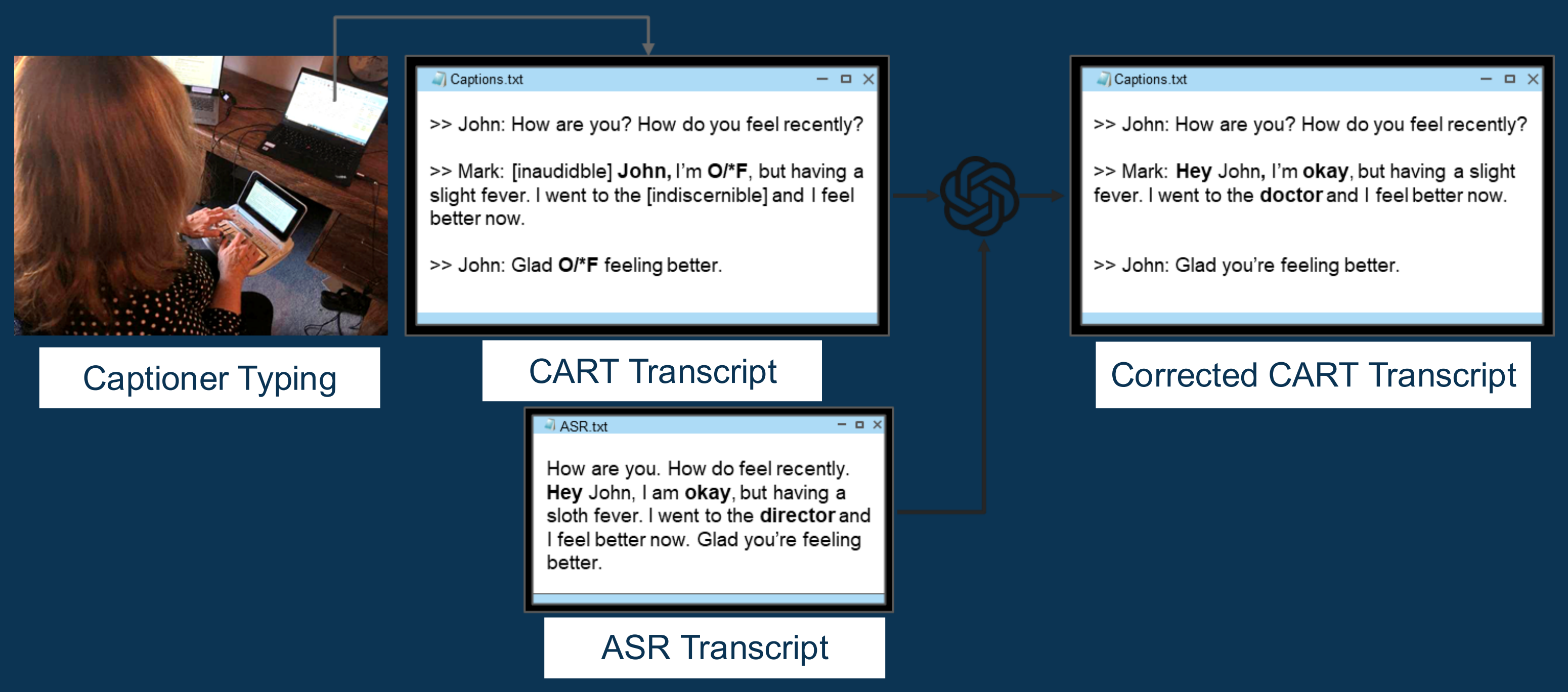
We interviewed with 10 professional CART captioners and concluded four categories of errors in CART captioning.

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

All captioners were supportive of the idea to reduce the errors with the help of computational technology.

CARTGPT

1. Search for errors
2. LLM generates replacement texts
3. Text post-processing



>> John: How are you? How do you feel recently?

>> Mark: [inaudible] **John,** I’m **O/*F**, but having a slight fever. I went to the [indiscernible] and I feel better now.

>> John: Glad **O/*F** feeling better.

How are you. How do feel recently. **Hey** John, I am **okay**, but having a sloth fever. I went to the **director** and I feel better now. Glad you’re feeling better.

>> John: How are you? How do you feel recently?

>> Mark: **Hey** John, I’m **okay**, but having a slight fever. I went to the **doctor** and I feel better now.

>> John: Glad you’re feeling better.

GPT-4

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}

Results

We collected four speech corpus with noisy real-world data and different conversation topics as our evaluation dataset. We individually sampled approximately 10 hours of speech from TED-LIUM, Patient-Physician Conversation, MIT OCW and CallHome.

We compare the word error rate (WER) of CART, ASR and our method, CARTGPT. CARTGPT has a significant 5.6% improvement over the origin CART outputs.

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

We performed initial user study with 3 DHH participants. After experienced with the traditional CART approach and our CARTGPT approach, we found:

- CARTGPT improved users’ comprehension
- No visibile time difference between CARTGPT and CART

Discussion and Future Work

We evaluated CARTGPT quantitively with a dataset and qualitatively with DHH people. We plan to recruit a total of 12 DHH participants. Also, we see several other directions in future work:

- CARTGPT with Audio Embeddings: Supplement the LLM with audio information to correct mistranslate errors.
- Human-in-the-Loop: 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 to be run on-device to enhance privacy.

Leo’s Website:

https://binomial14.github.io

COMPUTER SCIENCE AND ENGINEERING
UNIVERSITY OF MICHIGAN