Reference Resolution and New Entities in Exploratory Data Visualization: From Controlled to Unconstrained Interactions with a Conversational Assistant

Abari Bhattacharya[1], Abhinav Kumar[1], Barbara Di Eugenio[1], Roderick Tabalba[2], Jillian Aurisano[3], Veronica Grosso[1], Andrew Johnson[1], Jason Leigh[2] and Moira Zellner[4]

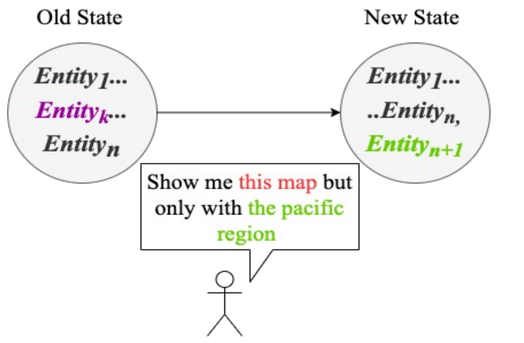
1. University of Illinois Chicago
2. University of Hawaii at Manoa
3. University of Cincinnati
4. Northeastern University

Contact:abhatt62@uic.edu

Our paper

This work is funded by awards 200757 and 2008986 from the National Science Foundation

Introduction: Reference Resolution and New Entity Creation



User's action: *Referring to an existing entity*
Reference Resolution: *Creating new entity (Entity $_{n+1}$) in response to users's action by modifying existing entity (Entity$_k$).*

Our broader goal..

- *Develop and deploy flexible conversational assistants to aid users explore data through visualizations*
- *Support user collaborations in exploratory data analysis*

In this work we..

- *Focus on reference resolution and new entity establishment*
- *Discuss reference detection and resolution in*
  - ✓ *controlled offline setting and*
  - ✓ *challenges presented when system is deployed "in the wild"*

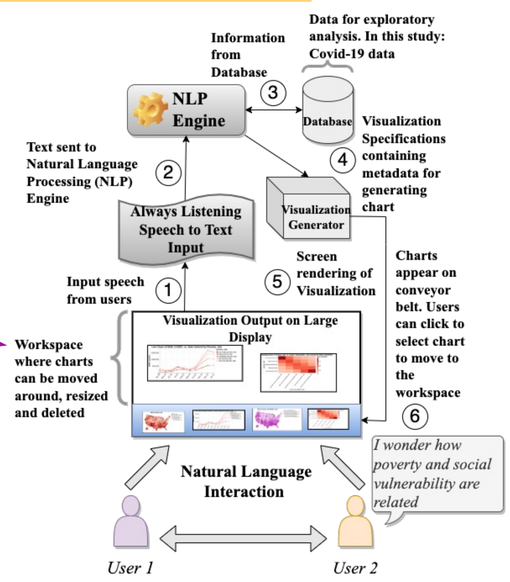
The "Articulate" Project: Background and Architecture

- *Data Collection for Chicago-Crime-Vis Corpus*
- *Multimodal System - Speech and Gesture*
- *Dialogue Management*
- *Reference Resolution Capabilities*

Articulate 2 Art-City-Asst → ***Articulate +*** → ***Articulate+ DM Art-COVID-Asst***

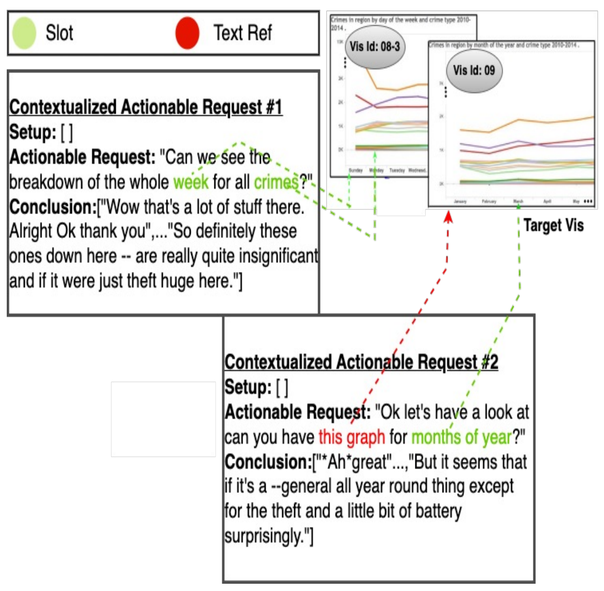
Articulate +
- *Overhearing*
- *Supporting Multi-user collaboration*
- *Only Speech*

Articulate+ DM Art-COVID-Asst
- *Articulate 2 ported to COVID Domain*
- *Dialogue Management*
- *Overhearing and Supporting Multi-user collaboration*
- *Only Speech*

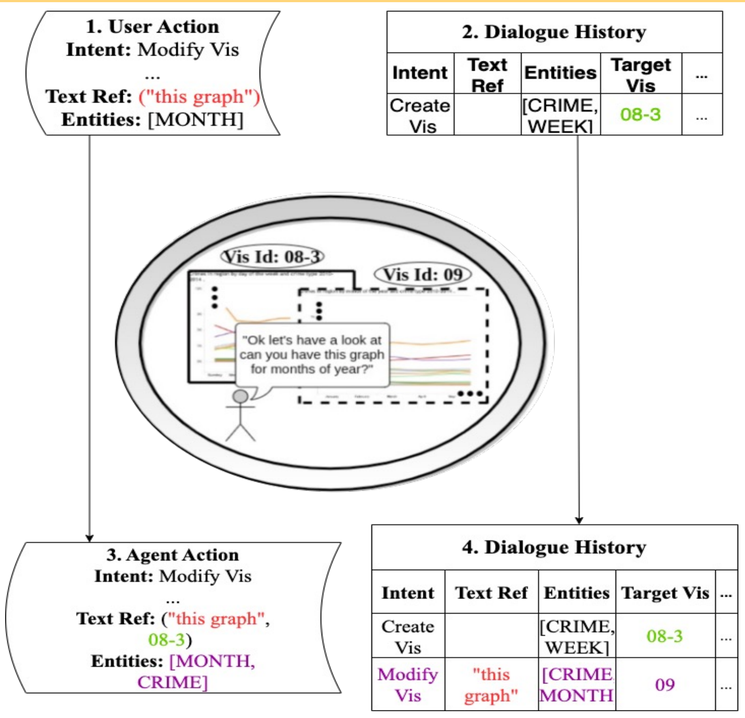


Corpus Creation: Chicago-Crime-Vis

User Study	• ***16 participants*** *interacting with human* ***Visualization Expert (VE)*** • ***3.2K Utterances***
Utterance Types	***Contextual Actionable Requests (CAR)*** • ***Setup:*** *Think aloud prior to an actionable request for VE* • ***AR:*** *The actionable request* • ***Conclusion:*** *Think aloud after AR.* • ***449 CARs covering 1545 utterances*** • ***AR annotated with Dialogue Acts (DA): Notably, CREATEVIS*** *(creating new visualization from scratch),* ***MODIFYVIS*** *(creating new visualization based on existing visualization)*
Referring Expression Annotation	***294 References****: 176 Textual* ***680 Phrases as slot fillers*** *corresponding to data attributes in the Knowledge Ontology* ***Inter-annotator Agreement:*** $\kappa = 0.85$
Chicago-Crime-Vis Knowledge Ontology	• *Constructed semi-automatically* • *3.5K terms categorized into 11 semantic pertaining to the city of Chicago*



Co-reference: Detection, Resolution and New Entity Establishment



Semantic Frame Construction (CREATEVIS and MODIFYVIS)	• ***"month of year" → MONTH in (1)***
Reference Detection	• *Sequence Tagging (IOB2 format)* • ***CRF Model with POS tags as features (F1 = 61.2% on B-REF, I-REF, O-REF task)*** • ***"this graph" in (1)***
Reference Resolution (Heuristics based on Similarity)	• *Candidate visualization with highest cosine similarity (threshold of 0.4,empirically established) selected.* • ***"08-3" in (2)***
New Entity Establishment (New visualization)	• ***new visualization is constructed ("09") with referent's frame representation to infer missing information ("08-3") → Agent action (3)*** • ***updated Dialogue History (4)***

User Study with Art-COVID-Asst

Porting to new domain: Art-COVID-Asst	• *Updating KO* • *DM, Reference Detection, Reference Resolution : same as Art-City-Asst*
User Study	• ***15 groups of 2*** • ***2 open-ended timed exploratory data analysis tasks*** *pertaining to COVID mitigation*
COVID Knowledge Ontology	• ***710 terms*** *categorized into* ***13 semantic slots*** *pertaining to COVID*

Controlled vs Unconstrained: Evaluation and Results

		Controlled Setup Art-City-Asst	Unconstrained Setup Art-Covid-Asst		Remarks
Evaluation Methodology		• *Offline evaluation run on transcripts of Chicago-Crime-Vis corpus* • *Focus on references occurring in setup and AR (specifically DA-s CREATEVIS and MODIFYVIS)* • *Focus on single referents and single targets*	• *Real-time evaluation* • *Speech recognition errors were a major bottleneck in the user study* • *Experiments with transcripts of the user studies generated using Whisper speech recognition model and fed to the back-end code →* ***COVID(T) – Transcripts (#utterances: 3096)*** • *Real time logs →* ***COVID(A)- Automatic (# utterances: 8440)***		• *For evaluation of Art-COVID-Asst we need to manually verify the results returned by the reference pipeline.* • ***A significant sample size is computed for both*** • ***Random significant sample of 340 (11%) utterances for COVID (A), and of 370 (4.38%) utterances for COVID (T).***
			COVID(A)	***COVID(T)***	
Semantic Frame Accuracy	***Correctly identified slots***	*% of Visualization Frames*			
	All	*55%*	*54%*	*52%*	*Visualization Frames for utterances with DA CREATEVIS and MODIFYVIS*
	At least 75%	*85%*	*60%*	*63%*	
	None	*7%*	*18%*	*16%*	
Reference Detection	***Accuracy***				
	Actionable Request	*55.0%*	*25.0%*	*45.8%*	*Five-fold cross validation accuracy reported for Art-City-Asst*
Reference Resolution	***Accuracy for Window sizes***				*No references resolved during the actual study*
	1	*74.4%*	*-*	*36.3%*	
	∞	*68.3%*	*-*	*54.0%*	

Discussion

Findings from Evaluation

Semantic Frame Accuracy	• *We report* ***partial accuracy*** *to provide* ***more nuanced analysis*** *of the assistant's performance* • *In dialogue-based application for data exploration like ours,* ***partially recognized VF can generate charts*** • ***This may help the users move forward.*** • *Irrespective of subpar performance of speech to text algorithm* ***more than 60% VFs had 75% or more slots correctly filled*** • *Attested by questionnaires filled by users post user study* ✓ ***Mean scores of 4 and 3*** *respectively for* ***usefulness of generated charts*** *and* ***ease of command system use*** *on a 5-point Likert scale*
Reference Detection	• *Unlike controlled study setting with one subject, when two people collaborate for exploratory task, three things happen.* 1. *They* ***talk to each other*** 2. ***Make requests*** *to the system* 3. *Finally* ***draw conclusions****.* • ***Reference detection*** *in real-time utterances extremely* ***complex****.* • *In the case of COVID (A), we also attribute the lack of accuracy to* ***speech-recognition errors****.*
Reference Resolution	• ***Speech recognition error*** *major roadblock for lack of resolved references in COVID(A)* • *In the unconstrained setting, we observe when two people are involved in the conversation, there are* ***more relevant entries in DH → expanding the window yields better results***

Future Work

Modeling user behavior for referring more distant visualization
Leverage multi-modality – gesture, eye gaze and head movement tracking, etc.
Experiments with Large Language Models