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

Anonymous ACL submission

Abstract

In the context of data visualization, as in other grounded settings, referents are created by the task the agents engage in and are salient because they belong to the shared physical setting. Our focus is on resolving references to visualizations on large displays; crucially, reference resolution is directly involved in the process of creating new entities, namely new visualizations. First, we developed a reference resolution model for a conversational assistant. We trained the assistant on controlled dialogues for data visualizations involving a single user. Second, we ported the conversational assistant including its reference resolution model to a different domain, supporting two users collaborating on a data exploration task. We explore how the new setting affects reference detection and resolution; we compare the performance in the controlled vs unconstrained setting, and discuss the general lessons that we draw from this adaptation.

1 Introduction

Conversation is understood in context. When the world, whether real or simulated, can change because of the user's actions, new entities are created by the processes that change the world itself: then, reference resolution, which links what the user refers to with objects in the world, is crucial for a dialogue system to effectively respond to the user, including to create new entities.

Our overall research program aims to develop and deploy flexible conversational assistants to support users, whether causal or professional, and whether alone or in teams, explore data via visualizations on large screen displays (large screen displays better support exploration and collaboration (Andrews et al., 2011; Rupprecht et al., 2019; Lischke et al., 2020)). In this paper, we focus on new entity establishment via reference in such contexts. We start from a corpus *City-Viz* in which a user exploring crime data in a large city interacts with a Visualization Expert (VE) whom they know to be a person generating visualizations on the screen remotely from a separate room. On the basis of *City-Viz*, we designed and developed a first version of our assistant which we call *City-Asst*. We will report the performance of *City-Asst* on reference resolution, and especially reference establishment, with respect to the transcribed and annotated *City-Viz* corpus, evaluated in an offline manner. The second part of our paper discusses the challenges that arose when we ported *City-Asst* to a new setting: two collaborators work together to assess COVID policies given geographic and demographic features of the data, and interact exclusively with the deployed *Covid-Asst* (see Figure 1). We will illustrate the many issues which degrade performance, from speech processing errors, to the adaptation of models to new domains, to the inherently more complex setting in which the assistant is now behaving like an overhearer of somebody else's conversations.

A disclaimer before we proceed: the purpose of this work was to adapt a previously developed conversational assistant and to evaluate it in a more unconstrained setting. We do not believe in chasing after the latest shiny approach, including ChatGPT[1], and undertake a potentially infinite loop of changes which would never bring us to real user studies. Additionally, we strongly believe in ecologically valid data, such as our *City-Viz* data. This data is by nature small, in fact tiny as compared

[1] https://openai.com/product/chatgpt

to most current datasets. We will return to these issues in the Conclusions.

2 Related Work

Conversational assistants for data visualization. Earlier work on conversational assistants for data visualization include (Cox et al., 2001), which established the benefits of using NL to generate visualizations for exploratory data analysis. In the ensuing 20 years, several such systems emerged in this area, see (Shen et al., 2023) for a systematic survey: e.g., DataTone (Gao et al., 2015a), Articulate (Sun et al., 2010; Kumar et al., 2016), FlowSense (Yu and Silva, 2020), Eviza (Setlur et al., 2016a) and DT2VIZ (Jiang et al., 2021). Along the lines of our work, (Tabalba et al., 2023) performed and evaluated user studies with conversational assistant for exploratory data visualization. However, unlike our system, lacks dialogue management as well as reference resolution. Lately, Large Language Models (LLMs) have started being integrated into visualization tools (e.g., PandasAI from OpenAI[2]), but not as part of a conversational assistant that keeps track of dialogue history.

Co-Reference Resolution as a field is as old and as vast as NLP; here we focus on its applications to visualization, which are hindered by several limitations: e.g., only referents to objects within the current visualization are handled (Sun et al., 2010; Gao et al., 2015b; Narechania et al., 2020), or only referents for follow-up queries on a current visualization are tracked (Reithinger et al., 2005; Setlur et al., 2016b; Hoque et al., 2017; Srinivasan and Stasko, 2017). As (Shen et al., 2023) concludes, "existing [approaches] mostly leverage NLP toolkits to perform co-reference resolution. Although useful, they lack detailed modeling of visualization elements" or, we would add, of what has transpired earlier in the dialogue. In contrast to this, we focus on reference resolution within an environment in which visualizations are dynamically added to and removed from the screen, and can subsequently be referred to. This requires *accommodating context change*, a notion first introduced by (Webber and Baldwin, 1992) in their discussion of new entities that are the results of physical processes as in cooking (e.g., *the dough* resulting from *mixing flour, butter and water*). In the 30 years since, not much work has been done on how to accommodate the creation of new entities (see (Wilson et al., 2016) for documents and (Li and Boyer, 2016) for tutoring dialogues about programming), and none in the visualization domain. Note we do not focus on multimodal reference resolution, another vast area (Navarretta, 2011; Qu and Chai, 2008; Eisenstein and Davis, 2006; Prasov and Chai, 2008; Iida et al., 2011; Kim et al., 2017; Sluÿters et al., 2022), even if we will briefly touch on deictic gestures in Section 3.

[2]https://www.kdnuggets.com/2023/05/pandas-ai-generative-ai-python-library.html

3 Controlled Dataset: City-Viz

Our *City-Viz* corpus comprises multimodal interaction for 16 subjects that explored public crime data in our city to better deploy police officers.[3] As noted, they spoke with a human VE who remotely created visualizations on a large screen, was not visible and did not speak back. The corpus contains 3.2K utterances. Since the user was encouraged to reason out loud about the patterns discovered via visualization, conversational turns often start with *think aloud*, followed by what we call an *actionable request* (AR) for the VE.

Using ANVIL (Kipp, 2001, 2014), we annotated 449 CARs (*contextual actionable requests*), covering 1545 utterances: a CAR consists of *setup*, i.e. think aloud prior to the AR (up to and including utterances that mention data attributes, if any); the AR; and the *conclusion*, the think aloud subsequent to the AR (also based on data-attribute mentions). While each AR is just one utterance, each of set-up and conclusion may include more than one (on average, 1.8 and 2 respectively). Figure 3 illustrates an excerpt from the corpus, comprising two CARs.

Each AR is annotated for user intent with one of 8 Dialogue Acts (DA) labels (with excellent intercoder agreement on the 8-way annotation, $k = 0.74$), including: WINMGMT for window management operations, e.g., closing, or minimizing; CREATEVIS for creating a new visualization from scratch; MODIFYVIS for creating a new visualization based on an existing one. The transcribed corpus is publicly available (URL withheld); and so is an augmented dataset built to alleviate data scarcity, comprising a 10-fold increase to 160 subjects covering approximately 15K utterances obtained via delexicalization and paraphrasing.

[3]We acknowledge that this task may be fraught in the era of Black Lives Matter in the United States. This data was collected prior to 2020, when the current awakening as concerns policing and racism surfaced to public consciousness.

Figure 1: User setting for COVID data exploration, with two collaborators

Referring Expression Annotation. We annotated both *text* (NPs) and *gestural* references to visualizations. Hand gestures were coded with various labels (e.g., the kind of gesture, the objects pointed to on the screen, and so on); approximately a third were identified as referential when they co-occur with text references. We labeled a total of 294 references in the 449 CAR's, of which 176 textual, and 118 gesture. We obtained an excellent intercoder agreement of $\kappa = 0.85$ with 2 judges on the full interaction from one subject. Given lack of space, and because in our unconstrained setting gestures were not relevant, we will not discuss gestures further. Within the 176 text references (of which 19 appear in set-up, 109 in AR, and 58 in conclusions), we also annotated 680 phrases as slot fillers corresponding to data attributes (i.e., *slots*) in our knowledge ontology (KO). The KO was semi-automatically constructed via external sources such as our city portal, augmented with synsets extracted from Wordnet[4] and Babelnet [5]; it comprises 3.5K total terms categorized into 11 parent types such as CRIME TYPE, NEIGHBORHOOD, TIME etc, of which about half are common nouns and about half proper nouns pertaining to our city.

[4]https://wordnet.princeton.edu/
[5]https://babelnet.org/

4 Co-Reference: Detection, Resolution, and New Entity Establishment

We briefly discuss the NLP engine (in the context of the full conversational assistant, see Figure 2), focusing on its reference resolution component. The NLU pipeline relies on an information state architecture with dialogue state tracking. After speech recognition (please see below for further discussion), traditional parsing and semantic role labeling are performed, and then a semantic frame is computed (see below). The dialogue management module is responsible for: classifying the intent of the user as one of the 8 DAs mentioned in Section 3; performing reference resolution; and updating and maintaining the dialogue history (DH). The NLP engine transforms the user request (when appropriate) into an SQL query; and in a visualization specification that is passed to Vega-lite2[6], a separate visualization interface software, to create a visualization of the data returned by the SQL query and add it to the display.

4.1 Semantic Frame Construction

Each time a visualization is mentioned in the dialogue (whether it refers to a previous one or not) our model looks for slots in the request to form its semantic frame. We find phrases that are in close proximity in the embedding vector space to terms in the KO, by using a domain targeted word embed-

[6]https://vega.github.io/vega-lite/

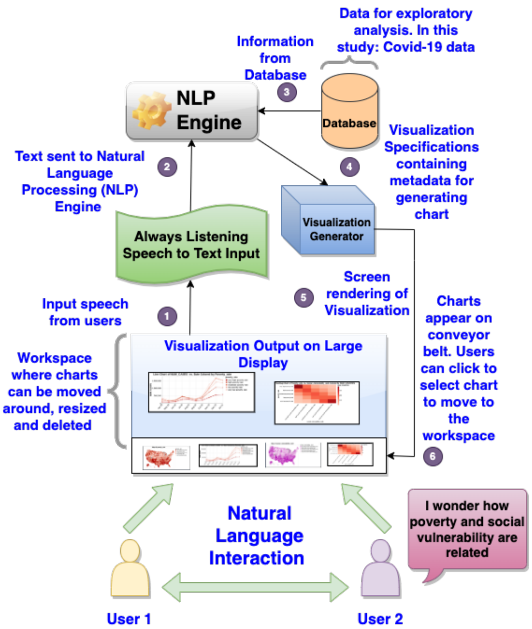

Figure 2: The Conversational Assistant (in its COVID incarnation, with two collaborators).

ding model (WEM)[7]. Subsequently the candidate words are pruned based on linguistic patterns using the SpaCy[8] dependency parse of the entire utterance to form the final list of slot fillers. For example in the AR in CAR #2 in Fig. 3, the prepositional phrase *"for months of year"* contains *"month"* and *"year"*, both of which are known as temporal slots in KO. Here, the terms are merged to form *"months of year"*, and mapped to the parent slot MONTH.

4.2 Dialogue Manager (DM)

The DM executes a dialogue policy which aside from making back-end decisions such as forming an SQL query for data retrieval, also seeks to populate unknown frame attribute values, via reference resolution if appropriate. For example, in Figure 3 the user asks to construct visualization *"09"* by using *"08-3"* as a reference. After processing CAR#1, the DH contains a single entry for *"08-3"*, and its specifications in a frame-slot format, including: the user intent (CREATEVIS), the type of plot, and its semantic frame in terms of attributes that were mentioned (*crime, week*). When AR#2 is processed and a CREATEVIS or MODIFYVIS DA is recognized, a new frame is created; while user intent (MODIFYVIS) and slots (MONTH, YEAR) are filled, others are left empty either because of under-specification by the user (e.g., axes labels, plot type, and so on) or they require additional processing by the DM.

When the semantic frame is complete, the state tracker adds it as a new entry in the DH while the system also outputs a json object (which we call a *visualization specification*) that instructs Vega-Lite2 to accordingly update the screen (i.e., add visualization *"09"* in this case); finally, the state tracker uses the DH to keep track of which visualizations are on the screen.

Reference resolution is carried out in two particular scenarios: a reference is detected and the user asks to (1) perform window management on a current visualization (e.g., close, move, and so on) or (2) create a new visualization based on an existing one (ie, a DA of CREATEVIS or MODIFYVIS is recognized). Next, we describe detection and resolution.

Reference Detection. We trained a sequence tagging model to detect text references (DTR). The model predicts tags using the standard IOB2 format (i.e., *"B-REF"/"I-REF"/"O-REF"* for beginning of / inside / outside text reference respectively). We trained a simple CRF model that uses POS tags as features, and two baseline models, BiLSTM-CRF and BERT-CRF. Further, to remedy data insufficiency - there are only 176 text references appearing across 449 CARs in the corpus, we investigated Sequential Transfer Task Learners and Multi Task Learners, in both cases, BiLSTM-CRF and BERT-CRF. As transfer or additional task, we use a NER task based on our augmented dataset, which is also automatically labelled for 23 NER tags, based on the B/I/O scheme: the *"B"* and *"I"* tag for each of the 11 parent slots in the KO (e.g., B-visualization, I-visualization) plus *"O"* tag (the slot names are known because they are manually labelled in the 449 CARs and delexicalization maintains their type).

Reference Resolution. To understand to which visualization the current referring expression refers, we use a mixture of recency and similarity. The slot fillers from the frame of the current referring expression and from the candidate visualizations in the DH are transformed into *visualization vectors*, ie, they are projected onto an embedding space along 11 dimensions, corresponding to the 11 slots in the KO, using the WEM mentioned earlier. Before comparing the two visualization vectors, a

[7] 100-dimensional continuous bag-of-word model trained on 5GB of online articles and wikipedia pages related to crime.

[8] http://spacy.io

recency factor is applied. If n represents the total entries in the DH, then the visualization vectors of the most recent $\frac{n}{2}$ entries in the DH are associated with a multiplicative factor of 1.0 signifying that they are equally preferred. The latter $\frac{n}{2}$ entries in the DH however are associated with a linear decrease by a factor of $\frac{1}{n}$. Finally, cosine similarity is used to score each visualization in the DH relative to the referring expression and the visualization with the highest score is selected, as long as it exceeds a cut-off of 0.40 (established empirically).

For the example in Figure 3, the DH contains only an entry for *"08-3"* (other earlier visualizations must have been closed and are not relevant any more). Since the cosine similarity score between *"08-3"* and the current semantic structure exceeds 0.4, *"08-3"* is chosen as the referent for *this graph*.

New Entity Establishment. Once the referent of the specific referring expression has been established, a new visualization (*"09"*) is constructed using the referent's frame representation to infer missing information (*"08"*). Explicit information in the current request is used to replace identical slots: e.g. MONTH, which was used to resolve the referring expression via WE embedding and cosine similarity among semantic structures, replaces WEEK as the temporal axis in *"09"*. Information that is unspecified in the request but present in the referred-to visualization is imported to establish the new visualization; in this particular case, CRIME is added to the slot list because *"08-3"* of the previous request includes it. Finally, plot type is set to line chart in the presence of temporal entities to better display trends across time. Otherwise, by default bar chart would be selected since 56% of all visualizations in the corpus are bar charts.

5 Constrained Evaluation on City-Viz

The results we present now were obtained by manually evaluating the pipeline, which was run on the transcribed *City-Viz* data in an offline manner: hence, we did not have to contend with speech errors, or with error propagation, since for every utterance, the DH up to that point was reset to a correct state if necessary. Currently, our model focuses on references occurring in *setup* and *AR* for detection, and in AR only for evaluation of semantic frame correctness. Additionally, we focus on single referents and single targets: e.g. in *"Can you bring up the graph behind the River North one?"*

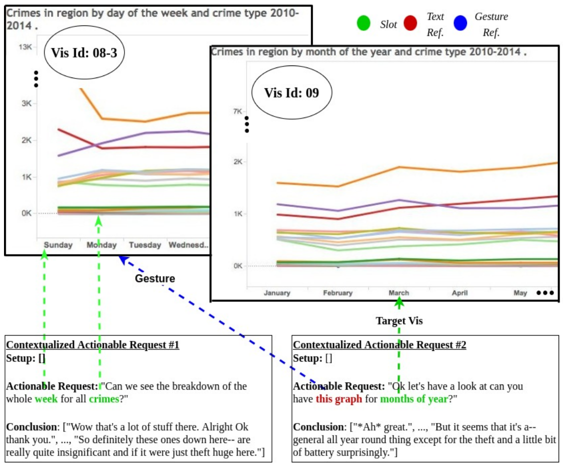

Figure 3: Excerpt comprising two CARs; references shown in red (text) and blue (gesture), and slot fillers in green. In CAR #1 visualization *"08-3"* is specified via temporal axis DAY associated with slot filler *"week"* and similarly CRIME for *"crimes"*. CAR #2 creates *"09"* substituting temporal axis DAY in *"08-3"* with MONTH, associated with slot filler *"month of year"* (identifiers are for clarity but not part of display).

Category	**Setup**	**Request**
Overall	19	109
Single Referents	18	86
Single Targets	14	66

Table 1: Text reference corpus counts.

the user refers to two visualizations; whereas *"well I would like to see battery by day of week, battery by month, and battery by year."* results in 3 new corresponding visualizations. However, our model only adds one of these visualizations to the dialogue history (DH) as part of the evaluation. Table 1 presents text reference counts only for setup and ARs (hence, excluding 58 references in conclusion). Single referents account for about 94.7% of references in setup and for about 80% of those in ARs. Finally, when filtering on single targets, we are left with the 80 text references (last row in table) on which we will focus.

Detection. Notwithstanding the lack of training data, the CRF performed the best, achieving a 61.2% F1 on the B-REF, I-REF, O-REF task. This is statistically significantly better than any other models (the next best is Multitask BERT-CRF with F1=43.5%). Hence, the CRF model is used in the subsequent steps in the pipeline. The five-fold cross validation accuracy of this CRF model on the City-Viz data is shown in Table 2.

Resolution. Accuracy on resolving text references for varying *WINDOW* sizes is shown in Table 3. If

	City-viz	Covid (A)	Covid (T)
Set-up	60.0	50.0	33.3
Request	55.0	25.0	45.8

Table 2: Evaluation of Reference detection model. City-viz: five-fold cross validation accuracy of CRF on City-Viz corpus; Covid (A): Accuracy of Reference detection model in real-time user study scenario; Covid (T): Accuracy of Reference detection model on correct transcripts of real-time user study scenario. Covid (A) and Covid (T) evaluated on a significant sample size

	Setup Window		AR Window	
	1	∞	1	∞
City-viz	85.3	85.3	<u>74.4</u>	68.3
Covid (T)	-	-	36.3	54.0

Table 3: Resolution accuracy for varying window sizes. Covid (T) evaluated on a significant sample size

one only takes into account the visualization introduced by the preceding AR (recall that we currently don't deal with multiple references), accuracy is 85.3% for set-up and 74.4% for AR. Interestingly, users also refer to the most recent visualization over 75% of the time. However, when we provide unlimited window size (∞ means all referent visualization candidates are eligible), resolution of references in ARs decreases; this suggests our linear decay function may need further tuning to better model the user preference behavior.

Semantic Frame Accuracy. While semantic structure construction is not used only for reference resolution, we report here its performance as concerns referring expressions in CREATEVIS and MODIFYVIS AR's. Our model achieved an overall accuracy score of 66.2% for semantic slots - this concerns the specification of the slots of the visualization frame in the DH, and includes slots that were explicit in the utterance, and those that were inferred. For example, given the example in Fig. 3, for "08" the two slot values are "crime" and "week", and for "09" "month" (explicit) and "crime", inferred via reference resolution. The first row of Table 4 shows that in the evaluation, 238 utterances have been recognized as CREATEVIS or MODIFYVIS; and provides the accuracy of slot recognition in quartiles, i.e. for 17 of these 238 ARs, no slots were correctly recognized, but for 131 (55%), all of them were correctly recognized; and in 83% of these ARs, at least 75% of the slots were correct.

	0%	25%	50%	75%	100%	Total
City-Viz	17	5	19	66	131	238
Covid (A)	22	1	25	8	66	122
Covid (T)	23	4	25	15	75	142

Table 4: Distribution of ARs wrt % correct slots. Covid (A) and Covid (T) evaluated on a significant sample size

	City-viz (H)	Covid (A)	Covid (T)
Set-up	218	73	149
Request	449	1296	2463

Table 5: Set-ups and requests in user studies (H: Human; A: Automatic; T: Transcript)

To generate the new visualization corresponding to a referring expression, the chart type (heat map, line chart, or bar graph) also needs to be inferred; it is simply copied from the referent.

6 Unconstrained setting: User studies in a COVID domain

A realistic evaluation of the NLP Architecture was conducted through user studies where two participants interact with the conversational assistant (*Covid-Asst*) to perform two open-ended exploratory data analysis tasks (inspired by (Tabalba et al., 2022, 2023)); the tasks are to uncover how factors such as access to doctors or elderly population may affect COVID mitigation strategies. Overall, 15 groups of 2 participants, performed the two tasks in a specified sequence, within a time limit of 25 minutes per task. We recruited the participants, aged 18+ from our University. With their consent, we audio and video recorded them, and collected logs generated by the back-end code of *Covid-Asst* for analysis purpose. As shown in Fig. 1, they are sitting and wearing a mike; also, each has a mouse with which they are able to reposition and click the visualizations on the screen. We encouraged the users to freely interact with each other and with *Covid-Asst*, and we did not provide specific instructions about the tasks, the interface, or the collaboration. The system is designed to "always listen" to the participants, whether or not they are addressing the assistant directly. This is implemented using the Web Speech API[9].

It was relatively simple to port *City-Asst* to *Covid-Asst* (Figure 2 shows the architecture) and mostly required to update the KO. For the Covid data, we identify 13 semantic slots like "Covid vulnerability rank", "Access to doctors", "Diabetes

[9] https://wicg.github.io/speech-api/

risk", “Uninsured rate" etc. and the possible values for these slots. As earlier, we enlarged the KO with synonyms for each slot and their values by using Wordnet and Babelnet to generate these synonyms. The generated KO has a vocabulary of 710 terms. This, as we describe in Section 4 forms the backbone of semantic slot filling and new entity establishment. We keep the same Dialogue Manager as before and use the best Reference Detection model built using the *City-Viz* corpus, namely, the CRF model. The Reference resolution algorithm also remains the same. Finally for screen rendering of the generated charts, the relevant data obtained from the database is converted to Vega-Lite grammar.

6.1 Findings of the User Study

To evaluate the reference detection and resolution pipeline in this setting, in principle we only need the log of the interactions to assess real-time performance wrt the utterances from the conversations of the participants. However, after we realized that speech recognition errors were a major bottleneck in the real-time study, we conducted additional experiments on the transcripts. These are generated using the Whisper speech recognition model [10] followed by light manual inspection. The corrected transcripts are then fed to the back-end code of the conversational assistant and new logs are generated. We name this version of the user study data as Covid (T) (for *Transcript*), while the real-time logs are named Covid (A) (for *Automatic*).

Since, as we noted earlier, reference detection applies to set-up and ARs, Table 5 shows the distribution of setups and requests in these two versions along with those from the City-Viz corpus. An important difference is that set-ups and ARs for City-Viz were manually annotated, whereas these are the results of automatic recognition for the COVID study (whether A or T). The table shows that there are many more set-ups in the City-Viz data; this difference is significant, as confirmed by $\chi^2 = 489.9511, p < 0.00001$ (with Bonferroni correction). There may be various reasons for this, one being that the classifiers that recognize setup and ARs were trained on the augmented City-Viz (Reference withheld) and perform worse here to start with. However, it is also possible that in fact, think aloud that feels natural when somebody is by themselves is not in a collaborative situation: a set-up by definition doesn't talk about a data attribute, but we surmise that the two collaborators are more focused on data attributes than on thinking aloud, precisely because they are interacting with another person.

For the purpose of the evaluation, we need to manually verify the results returned by the reference pipeline. Given the size of the data, we obtain two samples, one from Covid (A) (# utterances: 3096) and one from Covid (T) (# utterances: 8440). A significant sample size is computed for both with 95% confidence interval and 5% margin of error. This results in a random sample of 340 (11%) utterances for Covid (A), and of 370 (4.38%) utterances for Covid (T). Subsequently, we use Covid (A) and Covid (T) to refer to these samples of the respective groups, not to the whole group; all evaluation and analysis are done on these samples only.

6.1.1 Reference Detection

Table 2 shows the accuracy of the detected references in Set-up and Request utterances of Covid (A) and Covid (T). As expected, the performance degrades in a real-time user study scenario. Unlike the controlled study setting with one participant, when two people collaborate for an exploratory task, three things happen. First they talk to each other; next, they make requests to the system and finally they draw conclusions. These make reference detection in utterances extremely complex. In the case of Covid (A), we also attribute the lack of accuracy to speech-recognition errors.

6.1.2 Reference Resolution

We limit the evaluation of the reference resolution pipeline to Covid(T) as there were no references resolved during the actual study (note that useful visualizations may have been created all the same in response to those specific utterances, but not because a reference resolution was resolved). After conducting a thorough manual inspection of the issue we find the speech recognition errors to be the major road-block yet again. However using the corrected transcript (Covid (T)) we get a comparatively better performance as shown in Table 3. Since in this study setting only Actionable Requests (AR) where references are detected are resolved, we limit our evaluation to ARs only. Contrary to the constrained City-Viz setting, where considering only the previous AR was the better strategy, here

[10] https://github.com/openai/whisper (which became available on September 2022, after this conversational assistant was developed and hence could not be used for the user study)

limiting window size to 1 results in lower accuracy. We observe that in a more real scenario, especially when two people are involved in the conversation, there are more relevant entries in the dialogue history. This may also be due to the nature of the interaction with the large screen: in City-Viz, the user was standing in front of the large display, and often fairly close so that they would in fact mostly focus on only a portion of the display; in the Covid study, the two collaborators were sitting at about 6 ft from the screen (see Fig. 1), and hence all visualizations on the screen are more readily available to them.

6.1.3 Semantic Frame Accuracy

Similarly to what we observed in the controlled setting of City-Viz in Table 4, more than 50% AR utterances have all their semantic slots recognized as fully correct in the unconstrained settings with *Covid-Asst*. In fact, we see comparable performances of Covid (A) and Covid (T) across all quartiles. This shows that irrespective of the problematic performance of the speech-to-text algorithm, more than 60% ARs had 75% or more slots correctly filled and more than 80% ARs had at least 50% slots correctly identified. This also explains the reasonable success of the user study that we observed and is attested by questionnaires the users filled, despite the subpar performance of the speech-to-text algorithm.

7 Conclusions and Future Work

We have presented a reference resolution model for conversational assistants that help user in exploratory data visualization. In particular, the model resolves visualization references in the context of the current interaction, crucially tracking visualizations constantly being added to the screen. The model is central to the creation of new visualizations: visualization features encoded in the DH as slot values, help the model know how to refer to a visualization later on. We have also shown how the initial assistant, *City-Asst*, was ported to a completely different domain. We presented the evaluation of the reference pipeline in both settings, the constrained City-Viz and the "wild" Covid setting, in which two collaborators were exploring COVID data. We are fully aware that our results are not compared to an external baseline, but we contend that evaluations in grounded settings are important, and do not require creating some artificial baseline or evaluating the pipeline on existing reference resolution datasets.

Not surprisingly, the user evaluation brought several issues to the fore. First, we discovered that the speech API that we had chosen did not work very well (it would have been impossible to change it during the user study even if we had noticed it). Whereas this is unfortunate, we were able to obtain correct transcripts and run a second evaluation. Second, the nature of the interaction and the setting affected the conversations and the results: for example, we found many fewer set-ups in the COVID data, but on the other hand, more references to referents further back in the conversation.

Potential extensions for future work include ways to better model user behavior for referring to more distant visualizations and adapting our resolution algorithm beyond the cosine similarity measure to more sophisticated machine learning based approaches to take advantage of the rich visualization feature space in our case. Additionally, in the COVID user study users don't use hand gestures to interact with the screen, however they do use their mouses to click and reposition visualizations, hence bringing multimodality to the fore; not to mention gaze that can be approximated with head movement tracking, that another researcher in the group is investigating (see the instrumented caps in Figure 1).

Finally, as we had mentioned in the introduction, our goal was to evaluate our assistant in a realistic user study, and not jump into experiments with Large Language Models. However, we have started experiments in that respect, both as concerns the specific modules in our pipeline (for example, the embeddings of the semantic slots) and the system as a whole. So far, we have noticed that while ChatGPT (was released exactly after we finished the COVID user study) is able to generate charts in response to specific language instructions, if appropriately connected to visualization software, it is not able to resolve referring expressions, i.e., to create a new visualization whose specification is partly derived from the referent. But this will be the topic of a future paper.

References

Christopher Andrews, Alex Endert, Beth Yost, and Chris North. 2011. Information visualization on large, high-resolution displays: Issues, challenges, and opportunities. *Information Visualization*, 10(4):341–355.

Kenneth C. Cox, Rebecca E. Grinter, Stacie Hibino, Lalita Jategaonkar Jagadeesan, and David Mantilla. 2001. A multi-modal natural language interface to an information visualization environment. *International Journal of Speech Technology*, 4:297–314.

Jacob Eisenstein and Randall Davis. 2006. Gesture improves coreference resolution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*.

Tong Gao, Mira Dontcheva, Eytan Adar, Zhicheng Liu, and Karrie G. Karahalios. 2015a. Datatone: Managing ambiguity in natural language interfaces for data visualization. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software amp; Technology*, UIST '15, page 489–500, New York, NY, USA. Association for Computing Machinery.

Tong Gao, Mira Dontcheva, Eytan Adar, Zhicheng Liu, and Karrie G Karahalios. 2015b. Datatone: Managing ambiguity in natural language interfaces for data visualization. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, pages 489–500. ACM.

Enamul Hoque, Vidya Setlur, Melanie Tory, and Isaac Dykeman. 2017. Applying pragmatics principles for interaction with visual analytics. *IEEE transactions on visualization and computer graphics*, 24(1):309–318.

Ryu Iida, Masaaki Yasuhara, and Takenobu Tokunaga. 2011. Multi-modal reference resolution in situated dialogue by integrating linguistic and extra-linguistic clues. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 84–92.

Qi Jiang, Guodao Sun, Yue Dong, and Ronghua Liang. 2021. Dt2vis: A focus+ context answer generation system to facilitate visual exploration of tabular data. *IEEE Computer Graphics and Applications*, 41(5):45–56.

Hansol Kim, Kun Ha Suh, and Eui Chul Lee. 2017. Multi-modal user interface combining eye tracking and hand gesture recognition. *Journal on Multimodal User Interfaces*, 11(3):241–250.

Michael Kipp. 2001. Anvil-a generic annotation tool for multimodal dialogue. In *Seventh European Conference on Speech Communication and Technology*.

Michael Kipp. 2014. Anvil: The video annotation research tool. In Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, editors, *Handbook of Corpus Phonology*, pages 420–436. Oxford University Press.

Abhinav Kumar, Jillian Aurisano, Barbara Di Eugenio, Andrew Johnson, Alberto Gonzalez, and Jason Leigh. 2016. Towards a dialogue system that supports rich visualizations of data. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 304–309, Los Angeles. Association for Computational Linguistics.

Xiaolong Li and Kristy Boyer. 2016. Reference resolution in situated dialogue with learned semantics. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 329–338.

Lars Lischke, Lena Janietz, Anna Beham, Hartmut Bohnacker, Ulrich Schendzielorz, Albrecht Schmidt, and Paweł W Woźniak. 2020. Challenges in designing interfaces for large displays: the practitioners' point of view. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, pages 1–6.

Arpit Narechania, Arjun Srinivasan, and John Stasko. 2020. Nl4dv: A toolkit for generating analytic specifications for data visualization from natural language queries. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):369–379.

Costanza Navarretta. 2011. Anaphora and gestures in multimodal communication. In *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011), Faro, Portugal, Edicoes Colibri*, pages 171–181. Citeseer.

Zahar Prasov and Joyce Y Chai. 2008. What's in a gaze?: the role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 20–29. ACM.

Shaolin Qu and Joyce Y Chai. 2008. Beyond attention: the role of deictic gesture in intention recognition in multimodal conversational interfaces. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 237–246. ACM.

Norbert Reithinger, Dirk Fedeler, Ashwani Kumar, Christoph Lauer, Elsa Pecourt, and Laurent Romary. 2005. Miamm—a multimodal dialogue system using haptics. In *Advances in Natural Multimodal Dialogue Systems*, pages 307–332. Springer.

Franca Rupprecht, Carol Naranjo, Achim Ebert, Joseph Olakumni, and Bernd Hamann. 2019. When bigger is simply better after all: Natural and multi-modal interaction with large displays using a smartwatch. In *Proceedings of the Twelfth International Conference on Advances in Computer-Human Interactions (ACHI 2019)*.

Vidya Setlur, Sarah E. Battersby, Melanie Tory, Rich Gossweiler, and Angel X. Chang. 2016a. Eviza: A natural language interface for visual analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16, page 365–377, New York, NY, USA. Association for Computing Machinery.

Vidya Setlur, Sarah E Battersby, Melanie Tory, Rich Gossweiler, and Angel X Chang. 2016b. Eviza: A natural language interface for visual analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 365–377. ACM.

Leixian Shen, Enya Shen, Yuyu Luo, Xiaocong Yang, Xuming Hu, Xiongshuai Zhang, Zhiwei Tai, and Jianmin Wang. 2023. Towards natural language interfaces for data visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 29(6):3121–3144.

Arthur Sluÿters, Quentin Sellier, Jean Vanderdonckt, Vik Parthiban, and Pattie Maes. 2022. Consistent, continuous, and customizable mid-air gesture interaction for browsing multimedia objects on large displays. *International Journal of Human–Computer Interaction*, pages 1–32.

Arjun Srinivasan and John Stasko. 2017. Orko: Facilitating multimodal interaction for visual exploration and analysis of networks. *IEEE transactions on visualization and computer graphics*, 24(1):511–521.

Yiwen Sun, Jason Leigh, Andrew Johnson, and Sangyoon Lee. 2010. Articulate: A semi-automated model for translating natural language queries into meaningful visualizations. In *International Symposium on Smart Graphics*, pages 184–195. Springer.

Roderick S. Tabalba, Nurit Kirshenbaum, Jason Leigh, Abari Bhattacharya, Veronica Grosso, Barbara Di Eugenio, Andrew E. Johnson, and Moira Zellner. 2023. An investigation into an always listening interface to support data exploration. *Proceedings of the 28th International Conference on Intelligent User Interfaces*.

Roderick S. Tabalba, Nurit Kirshenbaum, Jason Leigh, Abari Bhattacharya, Andrew E. Johnson, Veronica Grosso, Barbara Maria Di Eugenio, and Moira Zellner. 2022. Articulate+ : An always-listening natural language interface for creating data visualizations. *Proceedings of the 4th Conference on Conversational User Interfaces*.

Bonnie Lynn Webber and Breck Baldwin. 1992. Accommodating context change. In *30th Annual Meeting of the Association for Computational Linguistics*, pages 96–103, Newark, Delaware, USA. Association for Computational Linguistics.

Shomir Wilson, Alan W Black, and Jon Oberlander. 2016. This table is different: A wordnet-based approach to identifying references to document entities. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 432–440.

B. Yu and C. T. Silva. 2020. Flowsense: A natural language interface for visual data exploration within a dataflow system. *IEEE Transactions on Visualization amp; Computer Graphics*, 26(01):1–11.