Conversational Localization: Indoor Human Localization through Intelligent Conversation

SMITHA SHESHADRI, Singapore Management University, Singapore
KOTARO HARA, Singapore Management University, Singapore

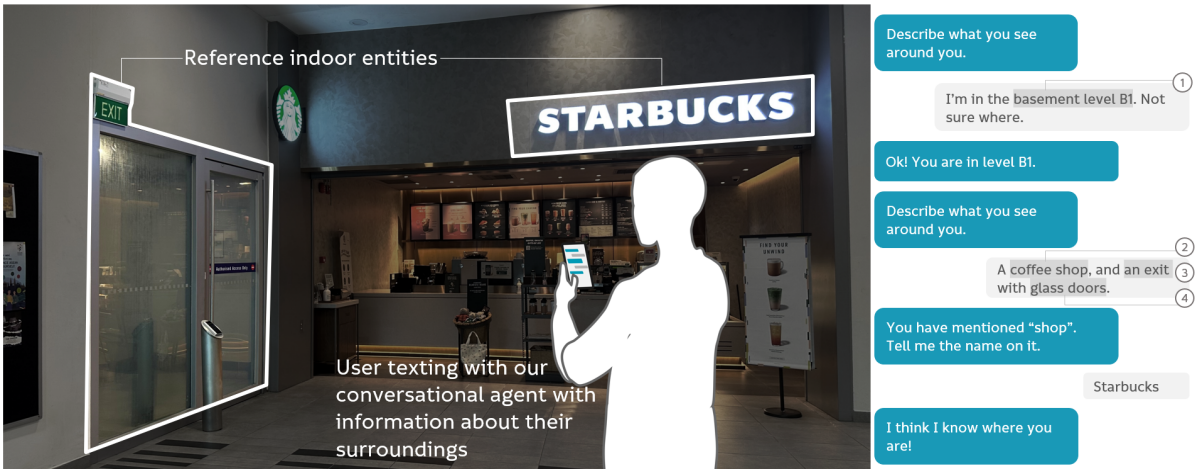

Fig. 1. A user who is unfamiliar with the indoor environment could interact with our conversational agent and supply information about their surroundings. Our system identifies localizationally valuable *entities* in the user utterances, compares them against a database of geographically indexed entities, and generates estimated location of the user. In this figure, the entity (1) conveys information about the level the user is in. The entity (2) has high potential to convey precise locational information, whereas entities (3) and (4) do not convey much locational information. Our system pursues entity "coffee shop" which it deems to be most useful.

We propose a novel sensorless approach to indoor localization by leveraging natural language conversations with users, which we call *conversational localization*. To show the feasibility of conversational localization, we develop a proof-of-concept system that guides users to describe their surroundings in a chat and estimates their position based on the information they provide. We devised a modular architecture for our system with four modules. First, we construct an entity database with available image-based floor maps. Second, we enable the dynamic identification and scoring of information provided by users through our utterance processing module. Then, we implement a conversational agent that can intelligently strategize and guide the interaction to elicit localizationally valuable information from users. Finally, we employ *visibility catchment area* and line-of-sight heuristics to generate spatial estimates for the user's location. We conduct two user studies in designing and testing the system. We collect 800 natural language descriptions of unfamiliar indoor spaces in an online crowdsourcing study to learn the feasibility of extracting localizationally useful entities from user utterances. We then conduct a field study

Authors' addresses: Smitha Sheshadri, smitha3shesh@gmail.com, Singapore Management University, Singapore, Singapore; Kotaro Hara, kotarohara@smu.edu.sg, Singapore Management University, Singapore, Singapore.

2474-9567/2023/12-ART176
https://doi.org/10.1145/3631404

with 10 participants at 10 locations to evaluate the feasibility and performance of conversational localization. The results show that conversational localization can achieve within-10 meter localization accuracy at eight out of the ten study sites, showing the technique's utility for classes of indoor location-based services.

CCS Concepts: • **Human-centered computing → Natural language interfaces**; *Empirical studies in HCI*.

Additional Key Words and Phrases: indoor human localization, conversational agent

ACM Reference Format:
Smitha Sheshadri and Kotaro Hara. 2023. Conversational Localization: Indoor Human Localization through Intelligent Conversation. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 4, Article 176 (December 2023), 32 pages. https://doi.org/10.1145/3631404

1 INTRODUCTION

GPS enables sufficiently precise outdoor localization for applications in navigation, tracking and monitoring [16], robot interaction [29], and mixed reality [30]. However, GPS is not suitable for indoor localization due to its low accuracy and reliability in enclosed spaces. This has led to numerous efforts to develop alternative methods for indoor localization, but a widely adopted solution has yet to emerge. Sensor-based localization solutions use a wide array of technologies including WiFi, RFID, or Bluetooth signals and localize targets (*e.g.*, a smartphone held by a user) [9]. However, these technologies face challenges, including deployment and maintenance overheads, reliability limitations, and lack of compatibility with existing devices [37]. Prior work has also investigated the feasibility of combining computer vision and user-captured images/videos to localize them. The computer vision-based solutions, however, suffer from viewpoint variations and noise, and often require extensive *a priori* data collection [9]. Jaisinghani *et al.* [20] detail the challenges in current indoor localization systems and highlight the need for a scalable, easy-to-deploy, and inexpensive solution.

We introduce *conversational localization*, a sensorless localization approach that leverages natural language conversation between a user and a conversational agent to estimate the user's location. In this approach, a user observes their surroundings and provides information that is useful for localization to the agent. The agent intelligently guides users to mention indoor entities that act as a reference for localization. Consider the following scenario:

Jon, a researcher, visits a collaborator's university for a meeting. Despite the array of signs, he finds himself lost in the network of interconnected buildings, corridors, and lecture halls. Though he finds a floor map, there is no "you are here' sign, leaving Jon uncertain about his exact location. He uses his phone and connects to the conversational localization service. The conversational agent asks Jon to describe his surroundings. Jon inputs things he can see around him, including " "corridor", "indoor plants", "staircase", and "lecture room". The agent asks for the number on the lecture room and Jon answers with "2 - 1". The agent processes this information and is able to identify Jon's location. It informs Jon of the building name, level, and the unique indoor entities around him. With this information, Jon identifies his position in the floor map and starts creating a mental path to his destination. (See additional scenarios in Appendix A.)

In implementing the prototype conversational localization system, we developed a custom named entity recognizer with grammar rules to extract entities from user utterances. To measure the suitability of the extracted entities for localization during a conversation, we devised an entity suitability scale, a semi-supervised method to rank the utility of entities. Our system estimates the user's position by generating candidate locations by crosschecking the extracted entities with a database of geographically indexed entity records. The candidates are transformed into an estimate by using Visibility Catchment Area (VCA) and line-of-sight heuristics. Our system deployment only requires the geographical database that can be constructed with image-based floor maps, albeit might require manual effort.

We conduct two user studies; first, to study people's language patterns in describing unfamiliar indoor environments and to evaluate our system's natural language processing prowess, we collected 800 descriptions from $N = 80$ participants through a study on Amazon Mechanical Turks. Second, we deployed our system on a university campus (spanning 13,00,00 m^2) and conducted a field study with $N = 10$ participants at ten study sites (spanning a total of 765 m^2). Our system achieved a within-10 meter mean accuracy at eight of the ten study sites, demonstrating the feasibility and promise of the conversational localization.

The main contributions of this work are:

- A novel conversation-based approach to indoor localization and development of a proof-of-concept system that implements the approach.
- An online crowdsourcing study that evaluates the feasibility of using natural language processing to extract localizational information from user utterances.
- Deployment of the system to our university campus and a field study with 10 users, which elicit benefits and challenges of the localization method.

2 RELATED WORK

2.1 Indoor Human Localization

Prior work has shown the efficacy of using sensors placed in the deployment sites to localize people in the indoor environments. For instance, the solutions that use WiFi-signals determine locations of user-held devices by measuring the strength and timing of the WiFi signals received with different routers (see [23, 37] for survey of WiFi-based localization solutions). WiFi-based localization, however, faces challenges with signal blockage and interference, and typically require a considerable number of routers for accurate localization in complex indoor locations [20]. Furthermore, fingerprinting of WiFi-based systems requires time, effort, and expertise. In their report of a 5-year experience in moving WiFi-based localization system from laboratory setting to real-world applications, Ni and Zhang *et al.* [26] report that for an average shopping mall, WiFi fingerprinting can take up to a week to collect, and an additional 1-5 weeks to stabilize.

Others have strategically placed Bluetooth Low Energy (BLE) beacons and determined the target's device location by measuring the strength and timing of the Bluetooth signals. BLE-based solutions face issues similar to WiFi-based solutions and, perhaps more critically, face limitations due to battery lives of the beacons [9]. RFID tags and readers can be used to determine the location of an RFID-tagged object by measuring the strength of the RFID signals received by the reader. Again, the solutions also face issues with signal blockage and interference and limitations due to line-of-sight requirements between the reader and the tag [9]. In addition to these technical challenges, the sensor-based methods require deployment of a signal sensor/emitter infrastructure and specialized expertise for setup and maintenance. Despite their benefits of high localization precision, sensor-based solutions are underutilized due to challenges encountered in the real-world environments. Some works make use of in-built sensors on mobile devices including accelerometer, compass, gyroscope to identify distinct patterns of movement such as climbing stairs, or taking escalators, as *semantic landmarks* [1, 33]. However, use of accelerometers often presents with lack of reliability [3] and gyroscopes are prone to abrupt measurement errors based on changes to the orientation of the device [1].

Gleason *et al.* [15] have investigated the efficacy of crowdsourcing infrastructure maintenance of localization systems to support indoor navigation. They highlight that the adoption of infrastructure-dependent localization solutions is hampered by the effort of installing and maintaining it over time. They propose to crowdsource long-term maintenance to non-experts. Though their approach could reduce the installation and maintenance overhead, its long-term viability is not clear. Thus, installment and maintenance overhead remains the burden for deploying these systems.

Another class of indoor localization methods uses images or videos captured by user's cameras to determine the location of the target. These solutions use image-processing techniques to compare the user-capture images or videos with a database of geo-tagged visual data to estimate the device's position within the indoor environment. For example, Gao *et al.* designed a method that used user-captured images of landmark objects (*e.g.*, signage) to triangulate the user's position [14]. Hybrid solutions that combine Wi-Fi data with images captured by users on smartphones have also been proposed [34]. Video-based solutions [22] use short videos captured by the user to identify reference locations for localization. Vision-based indoor localization solutions do not require the deployment and maintenance of complex hardware infrastructure like the aforementioned methods that relied on environmental sensors. However, these solutions often rely on large datasets of reference images or computationally-intensive techniques to extrapolate from limited reference images [9]. They can also be impacted by visual conditions such as lighting and image blur [19]. We aim to transfer the task of visual interpretation to users, thereby utilizing their inherent ability to understand spatial information. While prior work [4] ideates using gamification for situated data acquisition, our approach does not require on-site data collection or sensor deployment; we can remotely construct a database of geotagged indoor entities from floor maps.

2.2 Conversational and Natural Language Applications

Intelligent conversation has revolutionized how people interact with data and services. Conversational agents, more popularly known as chatbots, have grown increasingly prevalent [5]. The appeal of conversational agents lies in their ability to interpret natural human language to understand users' intentions. Practitioners in human-computer interaction foresee a transition from design-centric graphical user interfaces to a user-centric natural language-based interaction paradigm [12]. Researchers and designers have exploited the versatility of conversation for a whole array of applications. Acer *et al.* [2] studied the effectiveness of integrating natural language interaction into location-aware applications by devising a *hyperlocal* conversational agent for providing location-sensitive information.

In human navigation, conversation has been used as means for communicating automatically generated routing instructions to users. Duckham *et al.* [8] developed an outdoor landmark navigation model (OLMN) to generate landmark-based route instructions to support people's wayfinding activities in unfamiliar outdoor environments. Fellner *et al.* [10] adapted the model for indoor environments. Ohm *et al.* [27] studied visual saliency of various indoor objects using eye tracking. Furthermore, Fellner *et al.* [10] and Duckham *et al.*[8] supplied instructions to the users whose role is to passively follow the given navigational instructions. Ohm *et al.*[27] limit the identification of candidate landmarks to what is seen while following a predefined test route. These works recognize the importance of distinctive entities or "landmarks" to aid in human wayfinding. But these studies focus on the navigational activities after position is known. Wayfinding begins with obtaining the position of the object or subject [32]. While these works show the effectiveness of conversational agents in providing locational services and the localizational importance of landmarks, to our knowledge, no study has investigated the utility of intelligent conversation for indoor localization.

3 DESIGN PRINCIPLES AND REQUIREMENTS OF CONVERSATIONAL LOCALIZATION

We devised two design principles to create a conversational localization system by studying the issues with existing indoor localization systems. First, the system should be deployable without a bespoke sensing system. Thus, it could eliminate the deployment and maintenance overhead. Second, the system should not rely on a database of geotagged visual media that requires an *a priori* on-site survey. Instead, it should use data sources that can be obtained remotely (*e.g.*, floor maps that are publicly available online). By eliminating the need for an on-site survey, we intended to reduce the data collection overhead of our system.

In iteratively designing our system, we learned that implementation of a conversational localization system that meets the above principles is predicated on a user's ability to identify and convey information about their surroundings and the system's technical capabilities as we list below:

- **Human Ability to Identify Landmarks.** Our system should use information about indoor entities to localize the user's location. Our approach is different from the computer vision-based localization methods as it delegates the responsibility of identifying landmarks to the users. However, the feasibility of conversational localization is preconditioned on a person's ability to spot such entities in the indoor environment that is useful for localization.
- **Accurate Entity Extraction from Natural Language.** Indoor entities that the user identifies need to be communicated to our system. We use text-based conversation as means for communication. Thus, our system needs to accurately extract the entities that the user mentions in their utterances through natural language processing.
- **Reasoning about Localizational Utility of Entities.** Localizational value of information that the user provides varies; while an entity like a *"coffee shop"* that is locationally unique in the environment has high informational value, an entity such as *"table"*—an entity that can be seen everywhere on a university campus—has limited to no discriminating power to locate the user. Because the user could provide a mix of valuable and futile information, the system has to reason about what information to base location estimation on.
- **Agent's Capability to Navigate a Conversation.** Our system should take advantage of conversational mode to guide the users to mention information that could aid in localization.
- **Estimation of User Location based on Entity Information.** Our system needs to estimate the user's spatial position based on their proximity to mentioned indoor entities. While prior work achieved this using geographically-tagged datasets [22], we study alternative technique that leverage information available from only static sources such as floor maps to reduce data collection overhead.

We iteratively designed and implemented modules of our systems to meet these requirements. We conducted two studies to evaluate our system. First, to evaluate people's ability to mention indoor landmarks and technical feasibility of extract entities from user provided text and reason about their utility, we conducted an online crowdsourcing study (section 5). Next, to evaluate our system's on-ground performance, we deployed our system on our university campus and assessed its performance on localizing the user (section 6).

4 SYSTEM IMPLEMENTATION

Our implementation of the conversational localization system consists of four components: entity database, user utterance processing, conversational agent, and estimate generation.

4.1 Entity Database

Our system needs location data about the deployment site. The entity database contains information about the known indoor entities in the deployment site. To collect such entities, we manually annotated floor maps of our deployment site—*i.e.,* our university campus. Our university contains six interconnected buildings with five floors each. The whole of our university campus spans around 13,00,00 m² and is linked through an underground concourse. We had 34 individual floor plan images in the PNG format. We did not have access to spatial measurements of the floor maps.

4.1.1 Entity Taxonomy. In organizing the indoor entities in the floor maps, we define four entity subclasses to aid implementation of other system components.

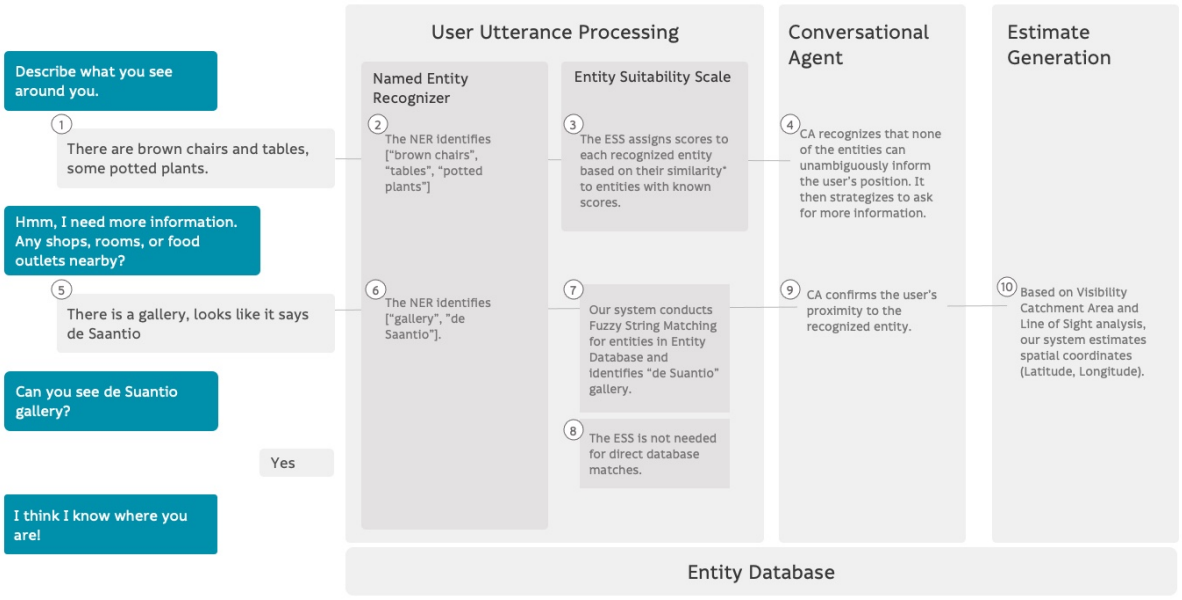

Fig. 2. A conversation taken from a user study log and the components of our conversational localization system. *Chat UI*: The user mentions entities in natural language through the chat UI (1&5) . *User Utterance Processing*: The named entity recognizer extracts entities from the user utterances (2&6). The system compare the the extracted entities with the entities stored in the *entity database* through exact or fuzzy match. When there is no match, the system ranks the potential utility of the extracted entities using *entity suitability scale* (3,7&8). Conversational Agent: Our system decides to either query more information (4&9). Estimate Genaration: The system generates location estimate and checks with the user if they are close to the estimated location (10).

- **Unique Entities:** Unique entities are those indoor entities that are unique to the given venue and have well-defined names to identify them (*e.g.*, restaurants' names). These entities' identities are defined with a unique string. Querying a unique entity eliminates all but one member of our entity database and thus provides highly valuable information for localization.
- **Partially-unique Entities:** A partially unique entity has aliases that are unique to a small handful of instances of similar entities. They typically have a numbering associated with them, such as "classroom B1-1" and "student study lounge L4-2." But because multiple buildings can have the same numbered room and partial string (*e.g.*, "classroom") is not sufficient to uniquely identify the location (*e.g.*, a user input "student lounge" yields many candidate entities), partially unique entities need more than just the name to be disambiguated.
- **Non-unique Entities:** Our floor maps also contain entities that typically do not have associated signage. They represent things like "staircases," "lifts," "escalators," and "washrooms." As there are several instances of such entities that cannot be disambiguated from one another without additional information, we regard them as non-unique entities.
- **Undesignated Entities:** In addition, we also anticipate that users will mention entities that are not on our floor maps. Such entities cannot be mapped directly to entities in our database. We deem these previously-unseen entities to be undesignated entities. To effectively handle interaction, our system needs to dynamically handle any new entities mentioned by users, which we explain in section 5.5.

Fig. 3. A sample floor map from our university that is used for the studies. Unique, partially-unique and non-unique entities as explained in section 4.1.1 are present around the floor map.

4.1.2 Data structure. The data structure we used to store entity details and an example entity is as shown in table 1. Every entity from the floor maps could be uniquely represented using the data structure. To obtain the geometry of the indoor entities from image-based floor maps, a member of the research team manually created polygonal geometries on top of the images and then used affine transforms to obtain the corresponding geographical coordinates. For certain indoor entities that were known with more names, they identified a list of aliases, i.e., the other names the entity was known by. The entire process of construction of the entity database from the 34 floor maps was completed in one day.

In total, there were 678 indoor entities in our floor maps, of which 96, 357, and 225 were unique, partially-unique, and non-unique entities respectively — see Appendix section B for sample list of indoor entities.

4.2 Utterance Processing

To identify and extract valuable information from raw user utterances, we devise two natural language processing modules—named entity recognizer and entity suitability scale—to handle raw user descriptions. The named entity recognizer receives raw user utterances and extracts all named entities that are potentially valuable for localization. For example, given the user utterance *"I am at the basement, near the escalator landing. I can see the coffee shop,"* it should identify *"basement level"*, *"escalator landing"* and *"coffee shop."* The entity suitability scale receives the extracted named entities and ranks their suitability for localization.

4.2.1 Named Entity Recognition. To identify locational information from user utterances, we develop a custom named entity recognizer (NER). The user utterances are pre-processed by removing punctuation and converting to lowercase. Then, a tokenizer breaks the utterance into composite tokens and a part-of-speech (POS) tagger

Table 1. A sample indoor entity data structure corresponding to an outlet of Subway (sandwich chain) present at Level 1 of a Building 1.

Entity Name	Subway
Entity Type	unique
Entity Aliases	[sandwich shop]
Entity Geometry	[[Lat-1, Long-1] ... [Lat-n, Long-n]]
Building	Building 1
Level	Level 1

tags the POS of the tokens. We used `regex_tokenizer` and `NLTK.tag.pos_tag` from Python's NLTK library [24]. We used our custom-defined grammar rules to match the tokens. We devised these rules by trial and error (see Appendix section C for the rules). We used `RegexParser` to compare and match the token sequences. The sequence matches identified by the RegexParser were the named entities.

We used fuzzy string matching to compare the entities on our database and the named entities extracted by the NER. A rigid string matching would not be able to identify entities with typographical or phraseological variations. Thus, we used the `FuzzyWuzzy` library in Python to conduct fuzzy matching. By trial-and-error, we set the fuzziness threshold to 0.70. If any of the named entities are a greater than or equal to 0.70 match to entities in our database or their aliases, we return these entities as matches. However, if the system does not find any direct matches from our entity database, it proceeds to analyze the entities recognized as undesignated entities and attempts to assess their suitability for localization through the entity suitability scale as described below.

4.2.2 Entity Suitability Scale. Indoor entities described in a conversation vary in their usefulness for localization. Consider the following description, *"To my right is a room [...] with glass doors. Behind me is a pillar in the middle of the hallway [...]"* (from a Section 5 study participant). The named entity recognizer extracts entities `["room"]`, `["glass" ,"doors"]`, `["pillar"]` and `["hallway"]`. As humans, we instinctually perceive a variability in the "usefulness" of the different entities. *E.g.,* the entity "room" would arguably be more useful in localization compared to "hallway". That is, we make a judgement about which pieces of information to use for localization and what to ask next in order to refine our estimation of the target's location. This indicates that (i) entities could be categorized into tiers based on their suitability for indoor localization, and (ii) an autonomous system should be able to rank the entities by their suitability so that it can utilize the best entities to localize people. We compiled the following four dimensions that characterize this "usefulness" of entities—which we call *Entity Suitability Scale*—by adapting existing literature in architecture, urban planning [25], indoor [10] and outdoor localization [8]:

i. *Signage-based imageability*: The visual characteristic of a useful entity should separate itself from its environment and let the user identify it with ease. Such a characteristic is referred as "imageability" [25] or "prominence" [10]; highly imageable entities possess distinguishable visual, cognitive, semantic, and structural elements [31]. Examples of prominent entities include signages like room number plates and restaurant signs.
ii. *Permanence*: A structure's permanence affect its suitability as a localizational cue. An entity which occupies the same space within a floor map for a reasonably long period of time can be considered as a permanent. Permanent entities contribute to localizational knowledge whereas temporary entities without a dedicated location do not [8].
iii. *Spatial Extent*: Narrower the area that the entity occupies, more precise the information it provides for locating a target. For example, a "shop" can contribute a more precise location than an entire "wall" [8, 10].

iv. *Ease of Mapping*: For an agent to utilize entities to localize a user, it has to know the entities' locations. That is, we need to be able to identify entities in data sources like floor maps and feed them into the system's locational database. Thus, we consider whether the entities can be found on floor maps as one of the dimensions.

We use this scale to investigate what entities people mention and how useful they are for indoor localization. Note, although some features may be more critical in assessing entities' suitability, we treat the importance of each feature equally. We come up with a if/then rule for each feature, and assign 1-point to an entity if it satisfies the condition (see Table 2). We then sum up the points to calculate the total suitability score of the indoor entity in question. A higher score on Entity Suitability Scale thereby indicates better suitability for indoor localization. While assignment of scores is manual, we enable the automated handling of previously unseen entities by obtaining their embeddings and finding similar known entities. We explain the details in sections 5.4 and 5.5.

4.3 Estimate Generation

In conversational localization, the user acts as the sensor of their environment. They mention the information they see in their surroundings which includes indoor entities. We use the mentioned indoor entities as points-of-reference and estimate the user's position by bounding their position to be in the visual vicinity of the mentioned entities. Upon receiving a set of one or more entities, this system component makes use of two heuristic techniques to generate spatial estimates; *Visibility Catchment Area* (VCA) and *Line-of-Sight tracking* (LoS).

4.3.1 Visibility Catchment Area (VCA) of entities. The area from which particular signage is visible is referred to as the "Visibility Catchment Area" of the entity [13, 35]. Visibility Catchment Area is calculated based on the type and size of the entity and the indoor geometry. For example, for signage of 152mm lettering, the maximum viewing distance is 30 meters. We first consider a circular VCA of 30-meter radius around the entity's geometric centroid for each entity. Then, we apply line-of-sight algorithm from the entity's geometry to trim the VCA and retain only those areas that can be viewed from the entity. Our reasoning is that if the viewer can see the entity, then the viewer is within that entity's VCA.

4.3.2 Line-of-Sight (LoS). Line-of-sight determines visibility of an entity by utilizing information about occlusions from other indoor entities and their bounding walls. We implement line-of-sight, and demonstrate the output in Fig.4. Under label 1 in Fig.4, the area in white corresponds to the VCA of entity UE1 after conducting LoS. By taking occlusions into acount of the entity visibility, we can make estimate of the location more precise.

Table 2. Entity Suitability Scale used to assess usefulness of indoor entities for localization. We take a sum of the scores assigned to each feature to compute entity's suitability score. (* A decision point is entity where a navigational decision can be made. *e.g.* Turn right at the ATM [10].)

Feature	Question	If yes,	Else,
Signage-based Imageability	Do instances of the entity in public places typically have signage?	1	0
Permanence	Are instances of the entity typically built into the infrastructure or connected to something which is built into the infrastructure?	1	0
Spatial extent	Is it a decision point?*	1	0
Ease of mapping	Is it on the available floor map?	1	0

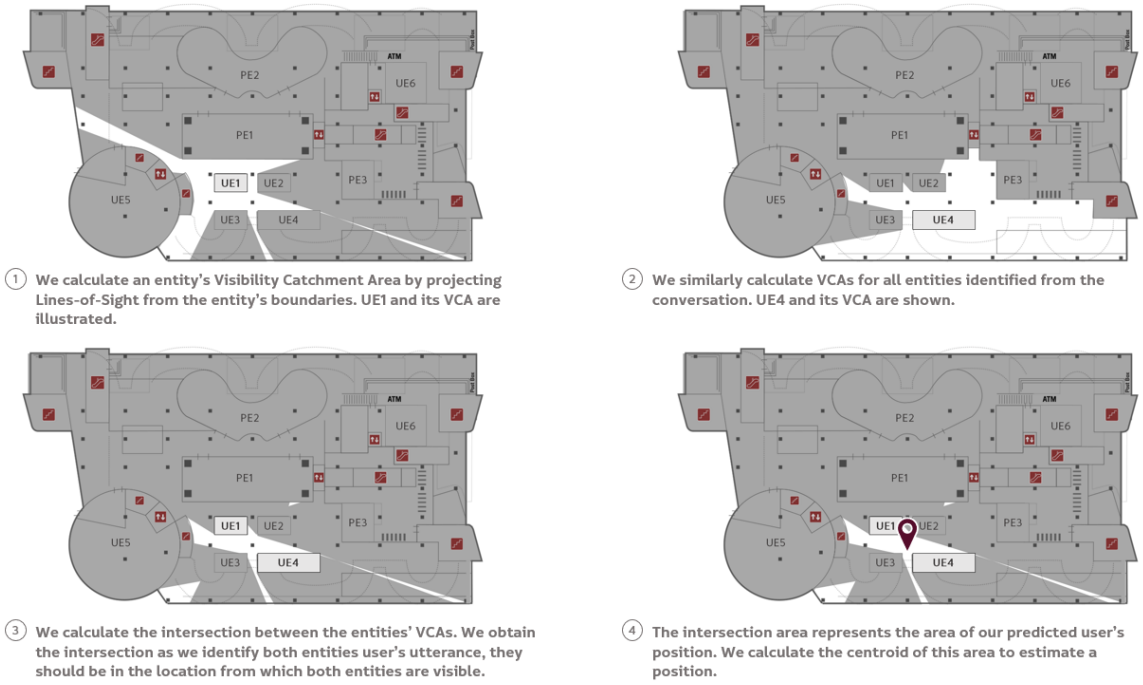

Fig. 4. Visibility Catchment Area (VCA) based user position estimation by applying a Line-of-Sight (LoS) algorithm. In (1) and (2), the lightened parts correspond to the area from which entity UE1 and UE4 are deemed to be visible according to VCAs. (3) When entities UE1 and UE4 are recognized in user utterance, the system intersects their VCAs. (4) The geometric centroid of the intersected area (red marker) is the estimated position.

4.3.3 Estimating the User Location. For each mentioned entity, we estimate the VCA using line-of-sight tracking. We then calculate the intersection of all the generated VCAs (see Fig.4 label 3). We regard the resulting area to be the our area estimate of the user's location i.e., if the user can see those entities, then the user must be in the visibility catchment area of all the entities. The estimated area of the user's position shrinks as and when more entities are mentioned, thereby allowing for a precise locational estimate. Once we obtain a polygonal intersection VCA, we calculate the geometrical centroid as the geographical estimate of the user's location. Although it is unlikely that the user is standing on the centroid, we use this point to enable the estimation of a point coordinate from a spatial area. Our system can thus use VCA and LoS to generate spatial coordinates by utilizing the entities mentioned to be around the user.

4.4 Conversational Agent and Chat User Interface

We design a chatbot user interface using HTML, JavaScript, CSS, and Python with Flask web framework (Fig.13). We deploy our agent on an AWS EC2 `t2.medium` instance. A user could access the system using a web browser. The design of the user interface is similar to many chat interfaces, where we have a conversation pane at the top and text field at the bottom. The user can type in their utterance into the text field. Clicking the 'send' button submits the utterance to the backend server and the message appears on the right side of the conversation pane.

Conversational agent's utterances appear on the left side of the conversation pane. In addition to plain text messages, it can show a selection and confirmation messages.

4.4.1 Dialog Acts. We define user and system dialog acts by following the CUED dialog act template [36]. We simplify the set of acts by retaining only those that are needed by our application and redefining them based on our need. The system and user dialog acts are listed in tables 3 and table 4 respectively.

4.4.2 Information Frame and Conversation Strategies. The agent guides the conversation to retrieve location information. It attempts to populate an information frame to fill four slots: building, level, entities, and entity matches. The building and level slots are populated once the agent recognizes building and level information, either directly from the user or through a confirmed entity. Entity matches contain entity information system recognizes in utterances. Once entity matches are disambiguated and confirmed, the entity slot is filled. This enables location mapping and estimate generation.

The agent attempts to populate the frame using five strategies that we have designed:

Strategy 1 – Identifying unique entities in user utterance: When user utterance contains unique entities, this strategy is possible. The agent passes raw user utterance to the user utterance processing unit, which returns the unique entity match. The agent populates the entity match slot and attempts to confirm that the user can see the identified match. Upon confirmation, the entity slot of the frame is filled, and the building and level information is identified from the confirmed entity.

Strategy 2 – Providing menu to select from candidate partially-unique entities: The agent attempts to identify partially-unique entities by presenting users with candidate matches when user utterance contains a partially-unique entity. The user can select one of the presented candidates or deny by ignoring them. If the user selects and confirms a candidate entity match, the agent populates the entity slot with the entity.

Strategy 3 – Applying constraint programming to disambiguate non-unique entities: When the agent identifies a partially-unique entity or non-unique in the user utterance, all instances of that entity are added to the entity matches slot. And when the user supplies additional information, such as the building name and level, the agent attempts to filter out the matches by applying constraints. If a sole match remains, the agent promotes the match to the entity slot. A sample interaction shown in Fig.5 (left).

Strategy 4 – Pursuing potentially valuable undesignated entities: User utterances may contain entities not on our floor maps. The agent decides the localizational value of the entities through the Entity Suitability Scale (Section 4.2.2). When the entities are deemed to have high localizational value, the agent pursues those entities by asking targeted questions and aims to employ one of the above strategies. A sample interaction shown in Fig.5 (middle).

Strategy 5 – Identifying and warning users about low-suitability, high-frequency entities: If the user keeps mentioning undesignated entities with low localizational information, the agent warns user about it and nudges them to mention high-valued entities. We determined classes of entities with low-suitability for localization from results of study 1 (See Section 5.4). A sample interaction shown in Fig.5 (right).

We point out here that the strategies can be applied only when corresponding entities in user utterances are identified. The strategies are attempted for each user utterance to attempt estimate generation (See appendix K for a sample interaction that uses strategies 3 and 4). In cases where none of the strategies are applicable, the agent simply requests for more description. We present a simplified flowchart that depicts the decisions that the system make in leading the conversation in Fig.6.

5 STUDY 1

Our conversational localization approach relies on the user's ability to identify and mention useful information from an unfamiliar indoor space, and our system's ability to extract relevant parts from user utterances. We

Table 3. List of dialog acts defined for agent utterance. Housekeeping dialog acts are excluded for brevity.

Dialog act	Definition	Sample utterance
`request()`	Request open description	Describe what you see around you.
`select(list(options))`	Get a choice from list of options from user.	Maybe you are around one of these entities. Try to select [Menu of options]
`guide(low_scoring entity)`	Present user with guidance.	You have mentioned “chairs”. I don’t have those on my floor maps. I only know about permanent places.
`confirm(entity)`	Get user confirmation about entity.	Confirm if you can see “classroom-1”:[Buttons for Yes and No]
`acknowledge(entity)`	Reiterate entity in user utterance	OK! You have mentioned "escalator"
`request_more(entity)`	Request more information about an undesignated high-scoring entity	You have mentioned "cafe". Tell me the name on it!

Table 4. List of dialog acts defined for user utterance with sample utterances from the study 1 (Section 5)

Dialog act	Definition	Sample utterance
`inform(unique entity)`	Informs agent about unique entity.	*A sign that says [shop name].*
`inform(partially-unique)`	Informs agent about partially-unique entity.	*Study lounge.*
`inform(non-unique entity)`	Informs agent about non-unique entity.	*I see escalators.*
`inform(undesignated entity)`	Informs agent about an entity that is not in the database.	*Chairs and table.*
`affirm(entity)/deny(entity)`	Response by selecting yes or no to agent’s confirm act respectively	(Selection of Yes/No button)
`choice(entity)`	Response by choosing from agent-provided menu to agent select() act	(Selection from menu)

conducted this study to 1. collect natural language descriptions of unfamiliar indoor places; 2. to inform the design of strategies by our conversational agent; and 3. to evaluate our system’s capabilities in processing raw utterances.

We collected 800 natural language descriptions of unfamiliar indoor environments from 80 participants. We found that our named entity recognizer (NER) and entity suitability scale (ESS) were effective in processing raw and unstructured descriptions and categorically identifying useful information for localization from them.

5.1 Method

We recruited $N = 80$ participants from Amazon Mechanical Turk (AMT) to describe indoor locations presented in 360° panoramic images. For the study, we captured 166 360° images from 24 locations across the university campus with an average of six images ($min = 5$, $max = 8$) per location. We selected ten locations from these 24 ensuring a combination of types of locations based on the entities present in them. The participants used a web browser interface to virtually look around the indoor environment of ten locations via pan-and-zoom interaction—see Appendix section E for the task interface. We asked the participants to assume they are describing a place they are currently present in. We note here that for this study, we did not use an interactive agent to collect interactions, instead we collected descriptions. This was because we needed to understand the language patterns

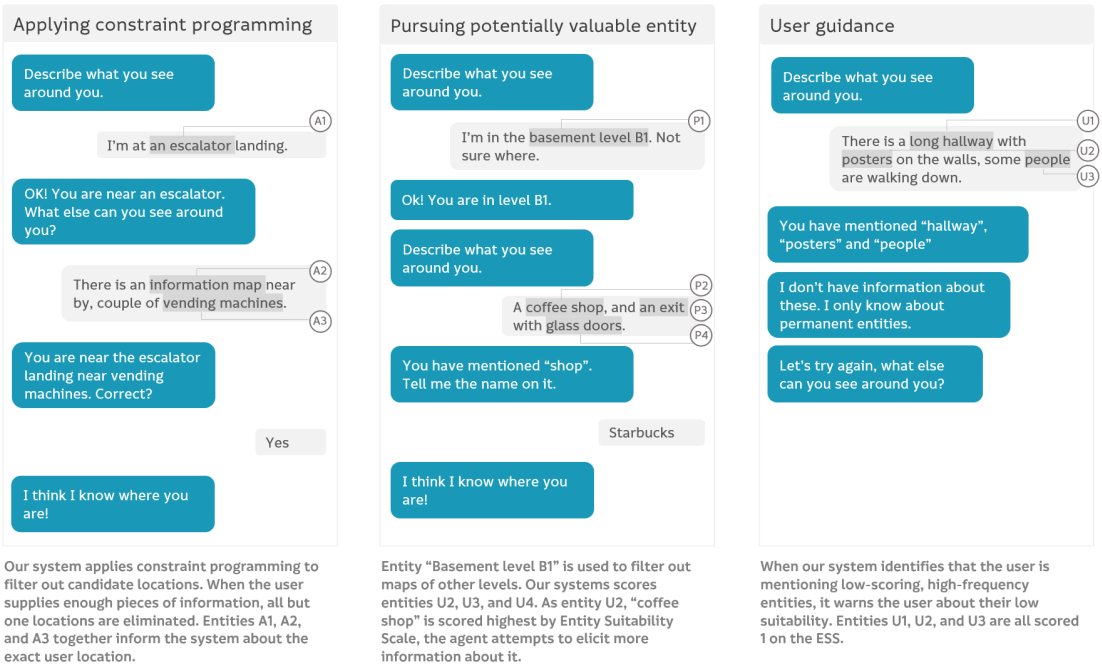

Fig. 5. We implemented strategies to guide users to mention information that is useful for localization as well as to make efficient use of mentioned information. As illustrated in these sample conversations, the conversational agent strategizes to arrive at an estimate by analysing entities in the users' utterances.

to devise the strategies employed by the conversational agent as explained in section 4.4.2. Chronologically, this study was conducted after we implemented the user utterance processing module and before the design of the conversational agent.

5.2 Result

Our floor maps contained 16 labels corresponding to entities in the location of the tours used for the AMT study. One research team member manually searched and identified 566 occurrences of the 16 entities in the 800 descriptions. This showed that our participants were capable of extracting useful entities from visual media.

5.3 Named Entity Recognition

Our NER extracted 99.11% (=561/566) of the manually identified entities (*i.e.,* recall = 0.9911). The NER missed five of the manually tagged entities, four of which contained typographical errors and one contained unusual phrasing (*"a restaurant that serves rice"* instead of naming the restaurant). Note that, we did not use an off-the-shelf entity tagger like SpaCy [18] as it missed many entities with proper nouns in their names (*recall* = 46.28% or 262/566). In total, including the 302 unique and 264 partially-unique entities, the NER extracted 4368 phrases, many of which are not on floor maps. We removed all duplicates, but still retained 2195 unique phrases. Among these 2195 unique phrases, as explained, 566 corresponded entities identifiable from the floor maps. The remaining 1629 entities corresponded to undesignated entities.

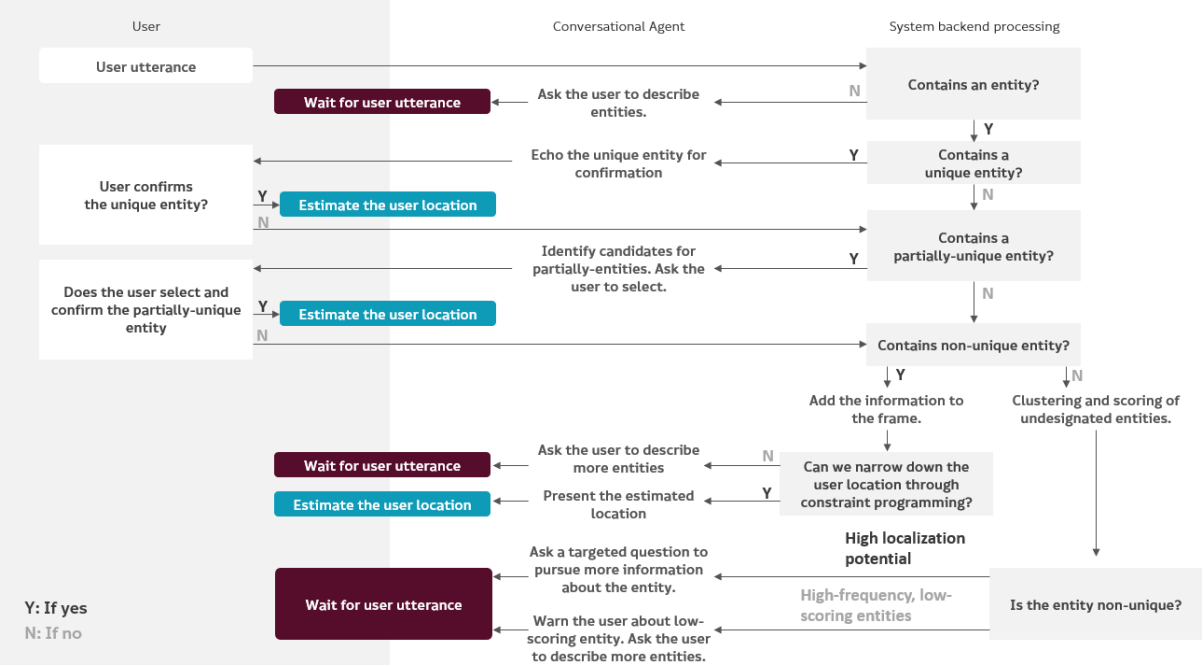

Fig. 6. A flow chart of a dialog adapted by our conversational agent. The actions and response of the conversational agent is determined by the type of entities in the user utterance, and previously mentioned information is the conversation history.

Out of the 800 descriptions, 167 contained at least one unique entity, 236 contained at least one partially-unique entity. 686 descriptions contained at least one undesignated entity. This indicates that while people identified entities useful for localization, they over-mentioned other entities too. This observation reinforces the need for our entity suitability scale that gives the system ability to rank entities by its localizational utility (section 4.2.2).

5.4 Entity Clustering and Evaluating Suitability for Localization

We used the following steps to create a model that ranks undesignated entities' suitability for localization in a semi-supervised manner. First, we convert each term into a vector of real values (*i.e.,* an embedding) using the Universal Sentence Encoder [6]. Encoding sequences of words into a vector of real values allow us to calculate similarity between two phrases. Second, we clustered the entities represented as embeddings using k-medoid algorithm. Through trial-and-error, we retained $k = 14$ clusters with 1806 unique entity phrases in total, *i.e.*, indoor entity classes. Finally, for each indoor entity class, one member of the research team manually calculated the suitability score following the scale specified in Entity Suitability Scale (See Appendix section D for the all 14 classes). As a result, each class had a score between 1 and 4, where 1 is less suitable and 4 is more suitable for localization. We conducted an analysis of frequency of mention of the entity classes and their suitability for localization (See Appendix section F for details). We found that certain types of entities that are mentioned frequently are low in their suitability for localization. We used this analysis to inform strategies used by the conversational agent (section 4.4.2).

5.5 Dynamic Handling of Undesignated Entities

The model allows us to assign suitability score to undesignated entities. We first encode the entity into a high-dimensional vector through the Universal Sentence Encoder (USE) [6]. Then, the embedding is compared against the 14 cluster centroids. If the undesignated entity's cosine distance to the cluster centroid is less than the defined threshold (= 0.75), we assign the suitability score of the cluster to the undesignated entity. Thus the system can handle the undesignated entity according to the assigned score. As an example, we make use of an utterance "I am near a food outlet, I can see some chairs and tables nearby." Our NER identifies ["food outlet", "chairs", "tables"] from the utterance. The fuzzy string matching does not yield any direct matches as the utterance did not contain entities from the database. The dynamic clustering and scoring recognizes that "food outlet" is similar to centroid "Restaurant/ Cafe/ Food Court" and assigns the corresponding score of 4 to "food outlet." Entities "chairs" and "tables" are assigned the centroid "Furniture" and the score of 1.

6 STUDY 2

6.1 Method

To evaluate our system, we deployed it to our university campus and we recruited $N = 10$ participants to visit ten study sites to perform conversational localization tasks.

6.1.1 Study Sites. Presence of types of entities and their prevalence would affect the effectiveness of conversational localization. For example, the presence of unique entities would reduce conversational localization difficulty. This was deemed to be the case because our participants in Study 1 easily identified unique entities.

To have a variety of types of entities at our study sites, we selected ten locations based on the types of entities in them (Table 5). We selected five locations that contained clearly visible unique entities. We anticipated that these five locations would enable easy interactions and quick estimate generations. In theory, sites 1 to 5 were localizable by applying strategy 1, i.e, identifying the unique entity. We selected the other five locations with no clearly visible unique entities in them. We deemed these five locations more demanding in terms of ease of interaction. Three of these five difficult locations contained partially-unique entities. That is, if a participant mentions a pair of the entity name and its uniquely identifiable number, we could identify the sole candidate of the entity and thus generate the accurate estimate of the participant's location. In this case, the agent would need to apply strategy 2, which is providing a list of candidate entities and asking the user the select the correct entity. Two of the five difficult sites contained non-unique entities. We selected these two locations so that if the participant mentioned a correct set of non-unique entities, we could narrow down the location of the participant. For these two locations, estimate generation would require strategy 3, or iteratively collecting enough information to apply constraint programming (section 4.4.2). Though varied in their task completion difficulties, all tasks could be completed. In the real world, this may not be the case as there are indoor locations without any notable entities. But a part of our goal was to assess if our conversational agent could guide the user to mention necessary entities for localization. Thus, we selected the sites so that they can complete a task in a reasonable number of conversational turns.

For each study site, we defined an enclosed walkable region that ranges from $52m^2$ to $103m^2$. We measure the distance between the centroid of this bounded region and the user's location estimated by our system to evaluate the localization error. We note here that since participants were free to move around within the site, the centroid would not exactly correspond to where our participant is standing at the instant of localization. But this way of measuring localization error is sufficiently precise to test the efficacy of locating the user within a multi-meter range. Alternatively, we could have instructed the user to stand in a much smaller region so the distance between the region's centroid and the location estimate generated by our system is a more precise depiction of the user localization error. But having the user in a very small region would overly restrict their

Table 5. Selected study sites and the types of entities in their vicinity. The area corresponds to the approximate area of the study site. Participants were free to move around within the boundaries of the study site.

Site number	Site difficulty	Unique Entity	Partially-unique Entity	Non-unique Entity	Area
1	Easy	Yes	Yes	No	103 m^2
2	Easy	Yes	No	No	70 m^2
3	Easy	Yes	No	Yes	119 m^2
4	Easy	Yes	No	Yes	52 m^2
5	Easy	Yes	No	Yes	52 m^2
6	Difficult	No	Yes	No	68 m^2
7	Difficult	No	Yes	No	72 m^2
8	Difficult	No	Yes	No	81 m^2
9	Difficult	No	No	Yes	94 m^2
10	Difficult	No	No	Yes	54 m^2

movements, which prevents us from observing how people would walk around and seek entities, and so is not a desirable study method.

Although we did not expect GPS signals to provide accurate location of the user in the indoor environment, we wanted to assess how better the conversational localization is relative to GPS. A member of the research team obtained the GPS readings by standing at the centroids of each study site using the same smartphone that our participants used.

6.2 Participants

We recruited 10 participants (4 female, mean age=30.2 years) who were unfamiliar with our university campus and had not visited the study sites before. We set this requirement to simulate the situation in which the user is visiting a new indoor location and trying to localize themselves. Herein, we will refer to the participants as P1 to P10. We used personal contacts and a word-of-mouth approach to recruit the participants. We reimbursed all our participants the equivalent of USD 21.75 upon completing the interaction at all ten study sites.

6.3 Procedure

Upon arrival on campus, the experimenter met the participant at a meeting point outside the ten chosen study sites. The experimenter briefed the participant about the study method, the conversational agent, and the tasks and obtained informed consent. The experimenter guided the participant to an example site and asked the participant to interact with the agent on the researcher-prepared smartphone to get acquainted with interacting with the agent.

The experimenter then guided the participants to the chosen study sites one after the other. We arranged the sequence of the study sites to minimize the travel distance. We did not change the order of the study site visits. This was done to fit the study duration within a reasonable time frame described in the informed consent. Once arrived at the study site, participants were informed that they were free to move around within the boundary indicators — see Fig.7.

Participants accessed the agent using a smartphone provided by us. We used a Xiaomi Mi A2 phone running Android OS. Participants accessed the conversational agent website using the Chrome browser. The experimenter opened a new session at each location and handed over the phone to the participants. We recorded the interaction transcript for all interactions.

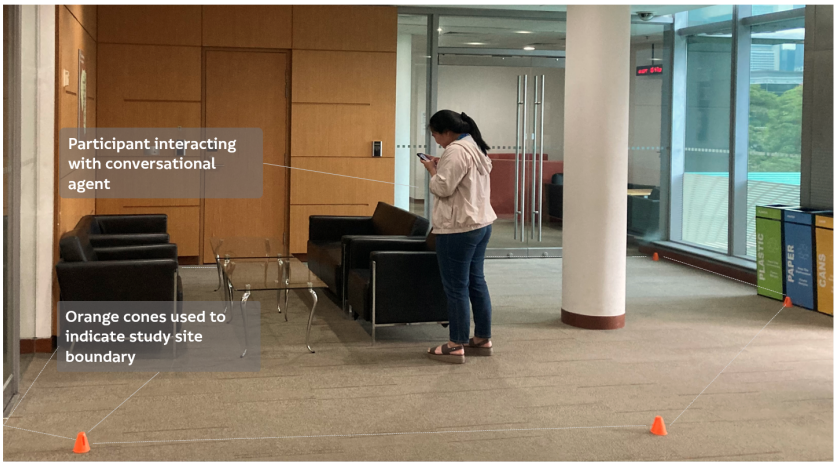

Fig. 7. A demonstrative image of a participant interacting with the conversational localization system during an on-site study session at site 10. The small orange cones indicated the boundary of the site.

6.3.1 Criterion for Termination. For evaluation, we identify three possible outcomes of each interaction. We regard these to be the three criterion for termination of interaction.

- **Correct location mapping:** When a user mentions an entity in their vicinity, and our system successfully identifies the entity, we consider it to be a correct location mapping. We note here that as long as the user and system identify an entity that the user is in the visual vicinity of, we categorize it as a correct location mapping, regardless of the distance between the user and that entity.
- **False location mapping:** This outcome occurs when either the user provides an entity that they are not in the vicinity of or when the system incorrectly identifies a false positive matching in the user utterance. We point here that for the location mapping to be carried out, the user needs to confirm their vicinity to the entity and respond to the system's `confirm` dialog act (Table 3). Therefore false mapping can only occur when the user mistakenly confirms an incorrect entity.
- **Conversation breakdown:** As shown in the dialog flow chart (Fig.6), our system resorts to asking for more information in the absence of any pursuable leads. To avoid indefinite conversation, we terminate the interaction between user and system after ten iterations. We define this outcome as conversation breakdown.

The system emitted the message *"Please inform experimenter"* after interaction termination, and participants were instructed to hand over the phone to the experimenter upon receiving that message.

6.4 Result

All 10 participants completed interactions at all 10 locations. Of the 100 interactions, our system generated locational estimates for 98 interactions. The other 2 were terminated by conversation breakdown. Our system kept track of the time duration participants spent for each interaction by calculating the time difference between their first and last utterances in that interaction. Participants took a mean duration of 59 seconds (min=7 seconds, max=306 seconds; Fig.8.a).

6.4.1 Localizational Accuracy. Of the 98 interactions with location estimates, 94 interactions were correct location mappings, 4 interactions were false location mapping. 82 interactions of the 94 with correct mapping had a

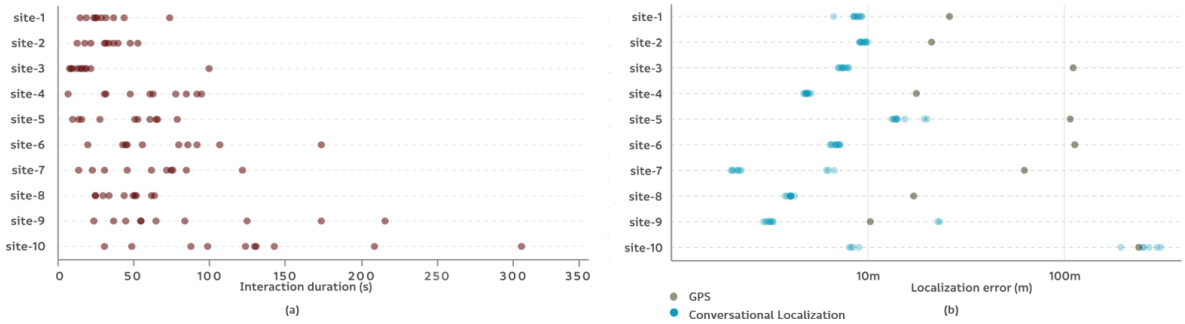

Fig. 8. (a)Time taken by participants at each site in seconds. In most interactions, participants spent less than a minute in identifying and supplying entity information. (b)Accuracy of estimates generated by the conversational localization system and GPS. Apart from the site 5 and 10, the system localized the users within 10m mean localization error.

localization accuracy within 10 meters, 10 interactions had a within 20 meter error and two interactions had 22 meter error from our ground truths. At 8 out of the 10 study sites, our system achieved a within-10 meter mean localization error. Site-5 had a mean error of 14 meter and site-10 had a 154 meter mean error.

6.4.2 False location mappings and conversational breakdown. Site-10 had the largest mean localization error of 154 meters. In site-10, the majority of participants (P1, P2, P3, P4, P5, P8, P9, P10) mentioned "Graduate Program Office," which was a signboard present on an office visible from within the study site. However, our university campus has two other Graduate Program Offices in different buildings. When our system identified that the user utterance contains "Graduate Program Office", it prompted the participants to select one of Graduate Program Offices. While the selection list also presented the building that each office is located in, two participants (P1 and P9) selected the wrong Graduate Program Office, which led to large localization errors. Our system failed in eliciting the required information from two interactions for site-10 (P2 and P4) and they engaged in long but futile conversations at site-10 where they mentioned a myriad of nearby objects (*"couches","trash cans", "glass windows"* by P2). These two interactions terminated in a conversation breakdown. The false location mapping and conversation breakdowns at site-10 caused the highly inaccurate estimate generation.

6.4.3 Correct location mapping with large errors. In site-5, a unique entity in the area (a shop) permanently closed during the period of this project. The entity name board were removed. As a result, our participants mentioned entities that were slightly farther away. Therefore, site-5 also witnessed a beyond-10-meter localization error ($mean = 14m$). For site-9, although the mean error was 7 meter, two interactions had (P1 and P7) had errors are 15 meters. There was a unique entity slightly outside the study site but with large and easily visible lettering. P1 and P7 mentioned this information and our system generated an estimate based on it.

6.4.4 Classification of localizational accuracy. Park and Lee [28] classify indoor positioning accuracy between 2-5 meters to be mid-high range accuracy, between 5-10 meters to be low-mid range accuracy, and errors larger than 10 meters as absence of any positional accuracy as they regard 10 meter errors to be too high for indoor positioning. On the same paper, they argue localization technology with < 10m resolution useful for applications like indoor work assistance and navigation (while they argued applications like path planning for people with disabilities would require 2-5 meters resolution). In 81 out of 100 interactions, we could locate a user within 10 meter error, and 48 were below 5 meter error. 19 interactions ended up in either conversation break down or location estimate above 10m. This analysis suggests that, for the majority of the times, conversational localization

provides sufficient indoor localization precision that is necessary for classes of location-based services like indoor work assistance.

6.4.5 Observations. P1 in site-7 mentioned that they were overwhelmed by all the signage around them. Site-7 consisted of student posters, and general direction board around the lounge door. On the door itself, signage about etiquette in the lounge and instructions to keep the lounge clean were present next to the lounge. P8 also brought up this issue and indicated that it would be helpful if the system guided users to locate and mention pertinent information. While we have included guidance about selecting the most appropriate type of entity, we have not added guidance about where to look for relevant information near each type of entity.

The next observation comes from P2's interaction in site-10. P2 saw and approached an information map in the area on which part of the floor map was provided for directional help. P2 mentioned that they wanted to mention entities present on the floor map but refrained from it as they reasoned that those entities might not be nearby. This reiterates the issue we recognized and tagged for future work from the user interaction study about disambiguating between signage on entities and signage on direction or information boards away from the entity.

During P3's study, the unique entity in site-5 was closed. The shutters also hid the name of the food court. P3 was able to find and mention the name of the food court from a nearby advert. This highlights a limitation to our approach posed by the indoor entities. Some locations in our database might be closed or occluded. We also observed some signages with partial legibility due to darkened or fallen-off letters. While this is not a system-induced limitation, we still keep note of it to address in our future work.

7 DISCUSSION

In this work, we proposed conversational localization—a sensorless indoor human localization method. We developed a prototype system that constitutes multiple system components to manifest the method. The system's conversational agent could interact with a user in natural language and elicit location entities that are localizationally valuable. The system extracted entities from the user's utterances with a custom named entity recognizer. Then, by making use of embeddings created by the universal sentence encoder, our system reasoned about extracted entities' suitability for localization. Finally, based on the extracted entities, the system could estimate the location of the user. Through two user studies, we evaluated the feasibility of accurately extracting entities that are useful for indoor localization from user utterances and assessed the efficacy of guiding a user to mention localizationally useful information in a conversation.

The result of the first study demonstrated that our system could extract 99.11% of the entities that the user mentioned. Our second study conducted on-site at our university campus showed that our system is capable of conversationally eliciting location entities from the users, and it demonstrated that the information is sufficient to locate the user with <10m resolution at eight out of ten study sites. Unsurprisingly, the localization resolution was lower compared to methods that rely on environmental sensors. However, this level of precision is already useful for implementing location-based services like indoor navigations [28].

7.1 Time-cost of Deploying a Conversational Localization System

The trade-off that we made in our design was the localization resolution and ease-of-deployment. Though lower in localization resolution, we argue that a conversational localization system is easier to deploy compared to a system that requires deployment of environmental sensors and fingerprinting. In order to deploy a conversational localization system at our university, for instance, we used floor maps like what we showed in Fig.3, a set of low-detail floor maps (the map with this level of detail is classified as level-of-detail 0 in Park *et al.* [28]), and then manually annotated and stored the entities therein. The process is not just more lightweight; such low-detail maps are often publicly accessible for communal venues like museums, conference centers, and airports. Thus, creating similar entity datasets for such venues is feasible even for those who have never physically visited on

site (*e.g.*, conference organizers). This is an important design advantage over the methods that require an on-site survey and fingerprinting.

To offer the informal evidence to show that it is easy to deploy our system, we invited a volunteer from outside of the research team to create the entity database—the vital system component that makes conversational localization work. The volunteer had no prior experience in localization research, thus deemed appropriate to represent non-expert individuals to make the system work. We asked the volunteer to use the same annotation tool our research team used to create the entity database. We first demonstrated how to use the tool to create entity geometries by drawing polygons on the floor maps and typing the names of the entities (see Appendix G for a screenshot of the tool). After a short practice session, the volunteer labeled 34 floor maps for our university campus. He annotated all 678 entities on the floor maps within 2 hours and 51 minutes. We visually inspected the created entities and found the only significant discrepancy between the volunteer-collected entities and the entities we created for our study was that the volunteer included one duplicate entity as he had annotated a particular staircase twice. Thus, the result suggested that a minimally trained person can provide the entity data that is necessary for conversational localization.

Requiring a few hours to collect data necessary for localization is arguably more lightweight and less labor intensive than deploying other solutions like WiFi and Bluetooth-based localization that requires fingerprinting—on-site process for registering sensor characteristics with ground truth location. To provide anecdotal evidence, we asked an engineer from another research group in our university working on WiFi-based indoor localization to perform fingerprinting for one of the floors of our campus. We shadowed the engineer as he fingerprinted one level of a building (see Appendix H for an image of the engineer collecting measurements). In total, the engineer took one hour to fingerprint the floor. In comparison, the volunteer we described above spent only 4 minutes and 26 seconds creating the entity database for the same area (corresponding to one of the 34 floor maps). The significant difference in two durations was not surprising because collecting entities was simple GUI-based annotation, while fingerprinting required walking to multiple points on the floor.

7.2 Interaction Design Challenges

Though our system could locate the user for the majority of the time, we observed interaction challenges that need to be addressed in the future. For instance, we observed through our study 2 participants' actions that users might not respond to the conversational agent's prompt accurately all the time. That is, when our system identified an entity in the user utterance and asked the user to confirm their vicinity to that entity, our participants confirmed even though they were not nearby that entity. This was problematic because then the subsequent estimation of the user location was based on inaccurate entities and their locations. We also noticed that users sometimes misclassified their locations when selecting from the drop down menu when the agent requested a response to `select()` dialog act (see Table 3). Again, such incorrect selection led to inaccurate user location estimation. The future work should investigate ways to make interaction more robust to such slips and mistakes. For instance, in our study 2, we saw false mappings as participants selected incorrect entities from list of candidates. We could provide additional information that can help participants differentiate between entities, perhaps by identifying unique characteristics of partially-unique entities.

Our study also revealed instances where conversational localization would fail. Entity misidentification by the utterance processing unit is a problematic but it is not a trivial issue to solve as we rely on the fuzziness of matching strings. Two strings which are similar in their spellings but represent different entities pose a challenge for our system. However, increasing the rigidity of the matching algorithm would result in more false negative errors as entities mentioned by the user might go unrecognized due to variations in spellings or phrasing. For our future, we aim to make entity identification more robust.

We observed that people could mention entities that are far away from the entity, such as what is on directional signs. Our participants had a good mental model of how the system works and knew entering information on a directional sign would result in inaccurate location estimation. Thus they did not enter such information. But this observation invites us to think about how to make a system that is robust to people entering entity names that are not actually near to the users.

Our conversation agent implementation completed interaction as soon as the estimate was generated. However, in practice, it needs to convey to the user the details of their location. Prior work states that people are not great at interpreting metric data [10], therefore we intend to research further into the type of information that users are likely to benefit from and that would help users gain more awareness of their surroundings. Furthermore, we are considering implementing an autocomplete feature for the messagebox in the agent so that the system can suggest matches for users. This is a non-trivial problem as it might lead to false mappings.

7.3 Incorporating Computer Vision into Conversational Localization

While we have shown that people can visually identify and communicate entities in the environment for our system to process, this approach could be less effective in some situations. For example, a user could be visually impaired and may not be able to identify entities in the environment. In such situations, it could be beneficial to allow the user capture images of their surroundings and supply it to the conversational agent. Then, the system can either (i) use Computer Vision (CV) techniques to elicit entity information that can be used for localization information from the images, or (ii) use CV techniques conduct localization by matching user-captured images to areas within the deployment.

To identify the challenges that might be involved in realising these two techniques, we (i) used object recognition and image captioning models to extract entity information from images, (ii) implemented an image matching algorithm that uses SIFT-based feature detection to classify images into location labels.

Object recognition and image captioning for entity identification. We captured five images in each of the ten study sites and ran them through object detection and image captioning models, in an attempt to elicit information similar to descriptions provided by users in conversational localization (See section 6 for details about the study sites). We used `YOLOv8` for object recognition and implemented it with the Ultralytics library in Python. Although the images included unique entities like Starbucks signage, the model failed to detect them. The entities that `YOLOv8` detected were non-unique and undesignated (*e.g.,* "couch", "dining table", "person").

We also used an image captioning model (`vit-gpt2-image-captioning`) to generate descriptions of the scene in a given picture, to investigate whether the generated descriptions included information useful for localization. We downloaded the model from HuggingFace [1]. However, the captions generated from the images looked like *"a room with a large glass"*, describing non-unique entities. Thus, it did not fare much better compared to `YOLOv8`-based entity recognition.

SIFT-based feature detection for location identification. We attempted to use an approach followed by prior work[21] that implements `SIFT`-based feature detection to conduct image matching. Approaches that use image-matching typically require a data-collection phase to capture images in all areas of the deployment site. The captured images are used as reference images and the created database is sometimes referred to as reference image database. We created a reference database using five images from each of our ten study sites by following procedure used by [21]; a member of the research team used an iPhone camera to capture images from around each study site by holding the camera at approximately the same height. Then, we captured five other images as query images in each of the study sites. Our approach used similarity score obtained by SIFT-based feature detection to assess the similarity between query image and reference images. To each query image, we assigned the location label of the study site associated with a reference image that was most similar to the query image.

[1] https://huggingface.co/nlpconnect/vit-gpt2-image-captioning

After following this procedure, we observed that 15 out of 50 query images were correctly assigned with the study site they were captured from. That is, localization was correctly carried out 30% of the time (see Appendix I for the confusion matrix). The classification accuracy was rather low compared to the the results from the prior work. This could be because of the lack of a sufficient number of reference images. For example, [21] use a reference image database of 60 images for a single office space.

Our results indicate that incorporating CV techniques into conversational localization is a potential topic for future work. In fact, research on using CV for indoor localization is an active research topic as we introduced in our related work (*e.g.,* [22]), where the existing work apply CV techniques like image matching [21**?**], object detection [7], and image captioning for localization [17]. While it would be useful for the user to have the option to use the camera input for identifying information, our cursory investigation suggests that more research and engineering are needed to create such a system, which can be used in situations where visual inspection of the environment is not possible, e.g., when the user is visually impaired.

7.4 Entity Data Population and Localizability

Some entities that our participants mentioned were not in our database because of their dynamic nature. Because a new business could emerge or an existing shop could close, floor map information can go outdated or the data on the maps could be obsolete. For example, twenty unique entities had changed over the 26 months of research duration on our university campus. Thus, if one wishes to keep the entity database consistently accurate, one must update the entity database about once a month. We expect, however, this number would change by factors like types of the environment. For instance, shopping malls may have frequent changes in tenant occupancy while museums have longer spans of fixed exhibition. We also observed that changes in unique entities were widespread during the last few years because many stores closed and new ones opened due to the Covid-19 lockdown.

More ephemeral entities like marketing adverts and maintenance signs would be even less likely to be on the floor maps, thus would not end up being in an entity database. Some of our participants mentioned such dynamic entities, assuming they are accurate and up-to-date, could improve localization capability of our system. Future work could investigate how to populate and update such information. Interesting future research directions include investigating the feasibility of asking our system's users to help populate the information; if the system successfully localizes the user, it could actively ask, *"do you see any other notable entities around you?"*. We should also investigate good user interface and interaction methods for situated end-users to annotate the indoor environment for populating and editing entity information.

Knowing how well a conversational localization system can identify a user's location in one facility would be useful in making informed judgment in deploying the system. If we quantify how well our system can localize a user at a given point in the facility, we could visualize such localizability scores across the facility. As the first attempt, we quantified the localizability by obtaining the number, type, and distance to the indoor entities in the vicinity. We calculate localizability for every pixel on our floor map and visualized the result in Fig.9. Though the formula that we used to calculate localizability is naive, this metric and its visualization already gives a sense of at which point at the indoor environment conversational localization could work well or fail. In Fig.9, the bright area is easily localizable and dark areas are hard. Future research could further perform simulation of localizability through conversational localization, as well as by combining multiple methods.

To estimate the portion of easily localizable area in our university campus, we calculated the spread of the VCAs of unique and partially-unique entities. Our reasoning was that if there is a unique or partially-unique entity visible from a place, users in that place might identify and mention that entity, thereby making it easily localizable. We found that 85% of our 13,00,00 m^2 university campus was covered by the VCA of at least one

unique or partially-unique entity. For our future work, we plan on devising a concrete heuristic for localizability to determine a location's suitability for conversational localization approach.

7.5 Incorporating Time-series Information to Improve Localization Quality

One of interesting future research directions would be, *"how can we combine conversational data and temporal information, such as estimation of how the user walked over time, to improve the localization quality?"* Consider, for example, our system conversationally locates a user at time t_0 as they mentions a unique entity. Then, after Δt seconds, the user mentions a non-unique entity at the new location. With the non-unique entity alone, our current implementation would fail to localize the user at $t_0 + \Delta t$. However, the fact that the system received information about the unique entity Δt seconds ago could be utilized to better estimate the user position.

To investigate the feasibility of using such time-series information, we simulated a user's movement within a floor of our campus by modelling human movement and using particle filtering. In constructing the movement model, we made a number of assumptions such as (i) a consistent walking speed of 1.42 meters/second (the average walking speed of humans), and (ii) user changed their direction uniformly at random when they hit a wall. In the simulation with this movement model, we assumed that the user mentioned a shop's name at time t_0. Then, after $\Delta t = 30$ seconds, user mentioned a "lift lobby," which is a non-unique entity [2]. At t_0, we create $N = 100$ particles within the catchment area of the unique entity. Then, particles move around following the motion model for Δt, dispersing themselves around the floor. At $t_0 + \Delta t$, the system receives the (simulated) user utterance "lift lobby", which could either correspond to lift lobby A and lift lobby B within the site. To assess which of the two lift lobbies is more probable based on what the user mentioned, we calculate distances between each particle and each lift lobby and the likelihood of each particle. The intuition is, more particles should be closer to the correct lift lobby given the motion model and the duration Δt, thus the likelihood is higher. Based on the computed likelihood, we resample the particles and update the distribution of the user position. As we showed in the figure in Appendix J, the particles more densely reside near the lift lobby A after the (simulated) user utterance of "lift lobby". Thus, we could predict that the user is near the lift lobby A. While this is a simple simulation, it suggests the feasibility of combining information collected via conversation and temporal data to disambiguating non-unique entities.

8 LIMITATIONS

We identified limitations in conversational localization. Its localization accuracy could be influenced by users' erroneous input. But we observed in the study that this does not happen frequently. We observed we can deploy conversational localization at 85% of the location in our university, but this may not be true for other buildings. The indoor environment has to be appropriate for locations with distribution of entities suitable for localization.

The system's localization accuracy cannot be sub-meter range, which an environmental sensor-based method could achieve. Our localization precision has a few meters of errors. But this should not significantly change our conclusion that conversational localization could be a viable option for building location services that require 10 m precision. We do not intend our approach to contend with existing solutions in accuracy of localization. Instead, we aim to showcase the feasibility, effectiveness, and ease of deployment of our solution.

Finally, we did not quantitatively evaluate the deployment through a ecological validity study. More rigorous testing with a deployed conversational localization system could shed light on challenges arising during use.

Additionally, we acknowledge that in the current form, our approach primarily relies on the user's ability to view and convey locational information. In the future, we aim to research and integrate methods to make our conversational localization accessible to people with blindness and visual impairments.

[2]To obtain an estimate for Δt, we asked a friend to walk from starting from the shop to the lift lobby and measured the time it took them to reach the lift lobby.

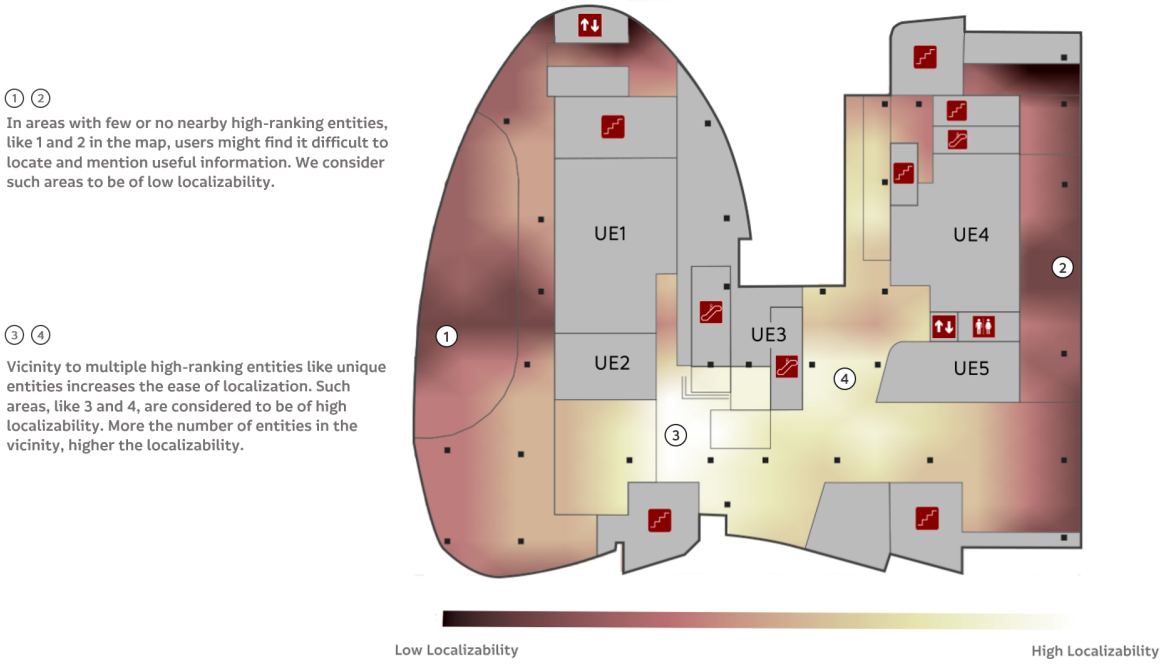

Fig. 9. A simulation of *localizability* of a floor in our university campus. Such simulation can help building management decide the appropriateness of conversational localization for their deployment usecase.

9 CONCLUSION

We ideated, designed, implemented, and evaluated a system to achieve indoor human localization through intelligent conversation. We constructed the required entity database by inspection of floor maps. We developed a custom Named-Entity-Recognizer with a recall of 99.11% on manually tagged indoor entities. We devised an application-specific Entity-Suitability-Scale to concretely and scalably determine suitability of indoor entities for localization. We evaluated the performance of these tools and studied user behavior when describing unfamiliar indoor environments by collecting 800 user descriptions from 80 participants. We applied VCA-based analysis to move from the language-based utterances to spatial estimates. We implemented a conversational agent with different strategies of processing to effectively guide users to mention high-valued information. We evaluated our system by conducting a study with 10 participants at 10 study sites spanning 765 m^2 in total. Our system achieved a within-10 meter localization accuracy in 8 of the 10 sites.

ACKNOWLEDGEMENTS

This research project is supported by the Ministry of Education, Singapore under its Academic Research Fund Tier 2 (Project ID: T2EP20220-0016). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore.

REFERENCES

[1] Heba Abdelnasser, Reham Mohamed, Ahmed Elgohary, Moustafa Farid Alzantot, He Wang, Souvik Sen, Romit Roy Choudhury, and Moustafa Youssef. 2015. SemanticSLAM: Using environment landmarks for unsupervised indoor localization. *IEEE Transactions on*

Mobile Computing 15, 7 (2015), 1770–1782.
[2] Utku Günay Acer, Marc van den Broeck, Chulhong Min, Mallesham Dasari, and Fahim Kawsar. 2022. The City as a Personal Assistant: Turning Urban Landmarks into Conversational Agents for Serving Hyper Local Information. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–31.
[3] Hoda Allahbakhshi, Lindsey Conrow, Babak Naimi, and Robert Weibel. 2020. Using accelerometer and GPS data for real-life physical activity type detection. *Sensors* 20, 3 (2020), 588.
[4] Joseph Berkner. 2019. Sensorless Indoor Localization Utilizing Collaborative Data Acquisition through Gamification (poster). In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. 556–557.
[5] Petter Bae Brandtzaeg and Asbjørn Følstad. 2017. Why people use chatbots. In *International conference on internet science*. Springer, 377–392.
[6] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175* (2018).
[7] Sukesh Davanthapuram, Xinrui Yu, and Jafar Saniie. 2021. Visually impaired indoor navigation using YOLO based object recognition, monocular depth estimation and binaural sounds. In *2021 IEEE International Conference on Electro Information Technology (EIT)*. IEEE, 173–177.
[8] Matt Duckham, Stephan Winter, and Michelle Robinson. 2010. Including landmarks in routing instructions. *Journal of Location Based Services* 4, 1 (2010), 28–52.
[9] Navid Fallah, Ilias Apostolopoulos, Kostas Bekris, and Eelke Folmer. 2013. Indoor human navigation systems: A survey. *Interacting with Computers* 25, 1 (2013), 21–33.
[10] Irene Fellner, Haosheng Huang, and Georg Gartner. 2017. "Turn left after the WC, and use the lift to go to the 2nd floor"—Generation of landmark-based route instructions for indoor navigation. *ISPRS International Journal of Geo-Information* 6, 6 (2017), 183.
[11]]ferbercombining Marvin Ferber, Mark Sastuba, Steve Grehl, and Bernhard Jung. [n. d.]. Combining SURF and SIFT for Challenging Indoor Localization using a Feature Cloud. ([n. d.]).
[12] Asbjørn Følstad and Petter Bae Brandtzæg. 2017. Chatbots and the new world of HCI. *interactions* 24, 4 (2017), 38–42.
[13] E Galea, L Filippidis, P Lawrence, and S Gwynne. 2001. *Visibility catchment area of exits and signs*. Vol. 2. Interscience Communications Ltd.
[14] Ruipeng Gao, Yang Tian, Fan Ye, Guojie Luo, Kaigui Bian, Yizhou Wang, Tao Wang, and Xiaoming Li. 2015. Sextant: Towards ubiquitous indoor localization service by photo-taking of the environment. *IEEE Transactions on Mobile Computing* 15, 2 (2015), 460–474.
[15] Cole Gleason, Dragan Ahmetovic, Saiph Savage, Carlos Toxtli, Carl Posthuma, Chieko Asakawa, Kris M Kitani, and Jeffrey P Bigham. 2018. Crowdsourcing the installation and maintenance of indoor localization infrastructure to support blind navigation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–25.
[16] Rajan Gupta, Manan Bedi, Prashi Goyal, Srishti Wadhera, and Vaishnavi Verma. 2020. Analysis of COVID-19 tracking tool in India: case study of Aarogya Setu mobile application. *Digital Government: Research and Practice* 1, 4 (2020), 1–8.
[17] Dhomas Hatta Fudholi, Abida N Nayoan, et al. 2022. The Role of Transformer-based Image Captioning for Indoor Environment Visual Understanding. *International Journal of Computing and Digital Systems* 12, 1 (2022), 479–488.
[18] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017). To appear.
[19] Junichi Ido, Yoshinao Shimizu, Yoshio Matsumoto, and Tsukasa Ogasawara. 2009. Indoor navigation for a humanoid robot using a view sequence. *The International Journal of Robotics Research* 28, 2 (2009), 315–325.
[20] Dheryta Jaisinghani, Rajesh Krishna Balan, Vinayak Naik, Archan Misra, and Youngki Lee. 2018. Experiences & challenges with server-side wifi indoor localization using existing infrastructure. In *Proceedings of the 15th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. 226–235.
[21] Hana Kubíčková, Karel Jedlička, Radek Fiala, and Daniel Beran. 2020. Indoor positioning using PnP problem on mobile phone images. *ISPRS International Journal of Geo-Information* 9, 6 (2020), 368.
[22] Mingkuan Li, Ning Liu, Qun Niu, Chang Liu, S-H Gary Chan, and Chengying Gao. 2018. SweepLoc: Automatic video-based indoor localization by camera sweeping. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–25.
[23] Fen Liu, Jing Liu, Yuqing Yin, Wenhan Wang, Donghai Hu, Pengpeng Chen, and Qiang Niu. 2020. Survey on WiFi-based indoor positioning techniques. *IET communications* 14, 9 (2020), 1372–1383.
[24] Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028* (2002).
[25] Kevin Lynch. 1960. *The image of the city*. Vol. 11. MIT press.
[26] Jiazhi Ni, Fusang Zhang, Jie Xiong, Qiang Huang, Zhaoxin Chang, Junqi Ma, BinBin Xie, Pengsen Wang, Guangyu Bian, Xin Li, et al. 2022. Experience: Pushing indoor localization from laboratory to the wild. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 147–157.
[27] Christina Ohm, Manuel Müller, Bernd Ludwig, and Stefan Bienk. 2014. Where is the landmark? Eye tracking studies in large-scale indoor environments. (2014).

[28] Junho Park and Jiyeong Lee. 2017. Establishing required LOD and positioning accuracy for indoor spatial information applications in public administrative works. *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography* 35, 2 (2017), 103–112.
[29] Robert Ross and Rahinul Hoque. 2019. Augmenting GPS with geolocated fiducials to improve accuracy for mobile robot applications. *Applied Sciences* 10, 1 (2019), 146.
[30] Saravjeet Singh, Jaiteg Singh, Babar Shah, Sukhjit Singh Sehra, and Farman Ali. 2022. Augmented Reality and GPS-Based Resource Efficient Navigation System for Outdoor Environments: Integrating Device Camera, Sensors, and Storage. *Sustainability* 14, 19 (2022), 12720.
[31] Molly E Sorrows and Stephen C Hirtle. 1999. The nature of landmarks for real and electronic spaces. In *International conference on spatial information theory*. Springer, 37–50.
[32] Ann Vanclooster, Nico Van de Weghe, and Philippe De Maeyer. 2016. Integrating indoor and outdoor spaces for pedestrian navigation guidance: A review. *Transactions in GIS* 20, 4 (2016), 491–525.
[33] He Wang, Souvik Sen, Ahmed Elgohary, Moustafa Farid, Moustafa Youssef, and Romit Roy Choudhury. 2012. No need to war-drive: Unsupervised indoor localization. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*. 197–210.
[34] Martin Werner, Moritz Kessel, and Chadly Marouane. 2011. Indoor positioning using smartphone camera. In *2011 international conference on indoor positioning and indoor navigation*. IEEE, 1–6.
[35] Hui Xie, Lazaros Filippidis, Steven Gwynne, Edwin R Galea, Darren Blackshields, and Peter J Lawrence. 2007. Signage legibility distances as a function of observation angle. *Journal of fire protection engineering* 17, 1 (2007), 41–64.
[36] Steve Young. 2007. CUED standard dialogue acts. *Report, Cambridge University Engineering Department, 14th October* 2007 (2007).
[37] Faheem Zafari, Athanasios Gkelias, and Kin K Leung. 2019. A survey of indoor localization systems and technologies. *IEEE Communications Surveys & Tutorials* 21, 3 (2019), 2568–2599.

A APPLICATION SCENARIOS

Conversational Localization provides a easy-to-deploy option for indoor localization. It appropriate for situations where the user can visually inspect and describe their surroundings and when accuracy of within-10 meters is acceptable. As the generated estimate of user's location can be obtained as coordinates relative to a specific floor map, the system can also be used as a localization module for navigational applications. We present below two application scenarios for conversational localization.

A.1 Use in Communicating One's Current Location using a Shared Vocabulary

Sara is exploring a new shopping mall. While walking along a long corridor, she notices that there is water spillage on the floor. She wants to alert the management of the shopping mall so that they can take measure to avoid injuries due to slipping on spilt water. However, as she is new to the mall, she is unsure of her location and needs help to convey the position of the water spillage to the mall management personnel. To find out her current location, she connects to conversational localization service. The conversational agent asks Sara to describe her surrounding. She describes a "long corridor", "some potted plants", "an escalator landing", and "some ATM machines". The conversational agent asks Sara to select the name of the bank providing the ATM machines from a drop-down list. Sara selects the appropriate bank name. The conversational agent is able to identify Sara's position and informs her that she is on "In corridor B, level 2 of building A, near [bank] ATM machines and exit 1 to subway station". Sara uses this information to alert the shopping mall management.

A.2 Use in Identifying One's Location for External Navigational Aids

Ben is at a new subway station, making his way to his platform. Ben tries to find directions to his platform but cannot find any helpful directional signages around him. To avoid any delays, Ben has downloaded a map of instructions. However, to use the map, he needs to figure out his current position so that he can start following the appropriate step in the instructions. Ben connects to the conversational localization service and inputs what he sees around him. His description includes "Starbucks coffee","Escalator landing", and "benches". Despite there being two Starbucks outlets in the subway station, the conversational agent is able to disambiguate between them based on information about escalators. The agent informs Ben that he is near "platform 4" and "platform

5" of the station. Ben is able to use this information to identify his current location and reaches his platform following instructions from his map.

B ENTITY DATABASE

Table 6. A partial list of indoor entites in our university campus. (gsr: group study room, sr: seminar room, cl: class room)

Entity name	Entity Geometry	Entity Type	Building	Level
escalator	[[l0,l'0],[l1,l'1]…[ln,l'n]]	non-unique	building 1	l1
escalator	[[l0,l'0],[l1,l'1]…[ln,l'n]]	non-unique	building 1	l1
lift	[[l0,l'0],[l1,l'1]…[ln,l'n]]	non-unique	building 1	l1
staircase	[[l0,l'0],[l1,l'1]…[ln,l'n]]	non-unique	building 1	l1
staircase	[[l0,l'0],[l1,l'1]…[ln,l'n]]	non-unique	building 2	l1
staircase	[[l0,l'0],[l1,l'1]…[ln,l'n]]	non-unique	building 2	l2
washroom	[[l0,l'0],[l1,l'1]…[ln,l'n]]	non-unique	building 2	l2
washroom	[[l0,l'0],[l1,l'1]…[ln,l'n]]	non-unique	building 1	l2
gsr 1-1	[[l0,l'0],[l1,l'1]…[ln,l'n]]	partially-unique	building 1	l1
gsr 1-1	[[l0,l'0],[l1,l'1]…[ln,l'n]]	partially-unique	building 2	l1
gsr 1-2	[[l0,l'0],[l1,l'1]…[ln,l'n]]	partially-unique	building 2	l1
sr 2-1	[[l0,l'0],[l1,l'1]…[ln,l'n]]	partially-unique	building 2	l2
sr 2-2	[[l0,l'0],[l1,l'1]…[ln,l'n]]	partially-unique	building 2	l2
cl 2-1	[[l0,l'0],[l1,l'1]…[ln,l'n]]	partially-unique	building 2	l2
gsr 3-1	[[l0,l'0],[l1,l'1]…[ln,l'n]]	partially-unique	building 4	l3
gsr 3-2	[[l0,l'0],[l1,l'1]…[ln,l'n]]	partially-unique	building 4	l3
starbucks	[[l0,l'0],[l1,l'1]…[ln,l'n]]	unique	building 1	l1
gym	[[l0,l'0],[l1,l'1]…[ln,l'n]]	unique	building 1	l1
subway	[[l0,l'0],[l1,l'1]…[ln,l'n]]	unique	building 1	l1

C NAMED ENTITY RECOGNITION: GRAMMAR RULES

Table 7. Rules of Grammar for named entity recognition.

Grammar rules	POS Tokens
NALL:{<NN|NNP|NNS|NNPS|JJ><CD|:>*}	CC :- Coordinating conjunction
NBAR:	CD :- Cardinal number
{<NALL.*>*<JJ.*>}	DT :- Determiner
{<NALL.*|JJ>*<NALL.*>}	IN :- Preposition or subordinating conjunction
	JJ :- Adjective
NP:	NN :- Noun, singular or mass
{<DT>*<NBAR><IN><DT>*<NBAR>}	NNS :- Noun, plural
{<DT>*<NBAR>}	NNP :- Proper noun, singular
{<DT>*<NBAR><CC><DT>*<NBAR>}	NNPS :- Proper noun, plural
	VB :- Verb, base form
	VBP :- Verb, non-3rd person singular present

D ENTITY SUITABILITY SCORES OF INDOOR ENTITY CLASSES

Table 8. Scores of Indoor Entity Classes obtained by clustering and manual scoring according to Entity Suitability Scale.

Indoor Entity Class	Signage	Permanence	Spatial extent	Ease of mapping	Total Score
Hallway/Corridor/Lobby	0	1	0	0	1
Floor	0	1	0	0	1
People	0	0	1	0	1
Furniture	0	0	1	0	1
Wall	0	1	0	0	1
Pillar/Column	0	1	1	0	2
Entrance/Exit	0	1	1	0	2
Staircase	0	1	1	1	3
Sign	1	1	1	0	3
Escalator/Elevator	0	1	1	1	3
ATM/Vending machine/ Information Kiosk	1	1	1	0	3
Restaurant/Café/Food Court	1	1	1	1	4
Seminar room/Office/Auditorium	1	1	1	1	4
Shop/Store	1	1	1	1	4

E STUDY 1 TASK INTERFACE

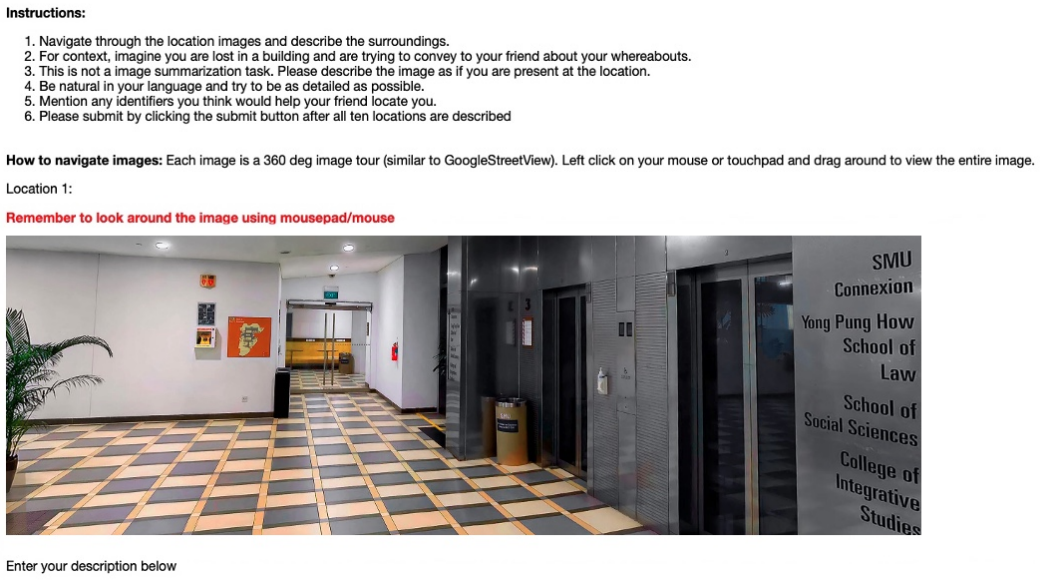

Fig. 10. Sample study layout of AMT study. Participants described the locations in text box provided below each location tour image.

F SUITABILITY OF ENTITY CLASSES AND THE FREQUENCY OF MENTION IN STUDY 1

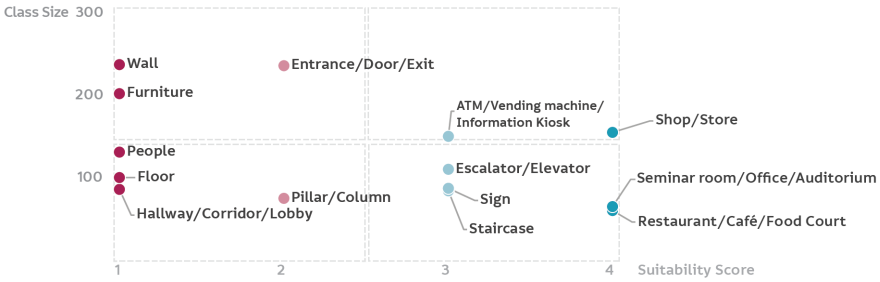

Fig. 11. A scatter plot showing entity suitability score-entity class sizes relationship—raw data in Appendix C. Entities in classes like "Shop/Store" are oft-mentioned and suitable for localization. Classes like "People" are rarely mentioned and not useful.

To understand the frequency of entities that people mention and how useful that information is in localization, we draw a scatter plot where each dot represents an indoor entity class (Fig.11); the x-axis represents the entity class's suitability score and the y-axis represents how often the participants mentioned entities in each class. We split the chart into four quadrants to make the following observations: (i) entity classes like "Shop/Store" and "ATM/Vending Machine" are often-mentioned and useful (top-right quadrant); (ii) entity classes like "Room" and "Restaurants" are rarely mentioned but would be useful for localization if mentioned (bottom-right), (iii) people often mentioned things like "Wall" and "Furniture", but they are less useful (top-left); and (iv) some entities are rarely mentioned and not useful (bottom-left).

G INTERNAL TOOL USED FOR ANNOTATING ENTITIES IN FLOOR MAPS TO CREATE THE ENTITY DATABASE

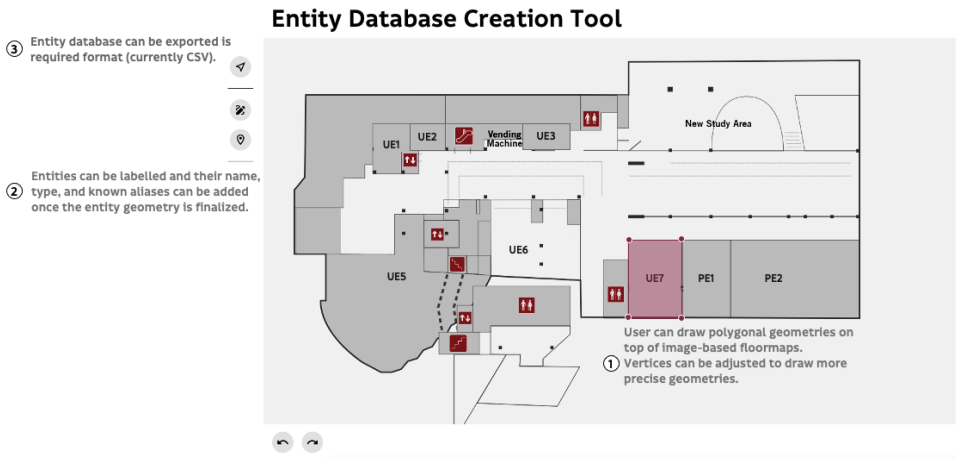

Fig. 12. The internal tool (developed using HTML/JavaScript) used to create entity geometries on image-based floor maps and to populate the entity database. The annotator can draw polygons by clicking on the vertices of the entity. The tool then allows the annotator to type the name of the entity by providing a text box. The entities can be exported into a CSV format.

H SHADOWING A RESEARCH ENGINEER WHO WAS COLLECTING DATA FOR A WI-FI-BASED INDOOR LOCALIZATION SYSTEM

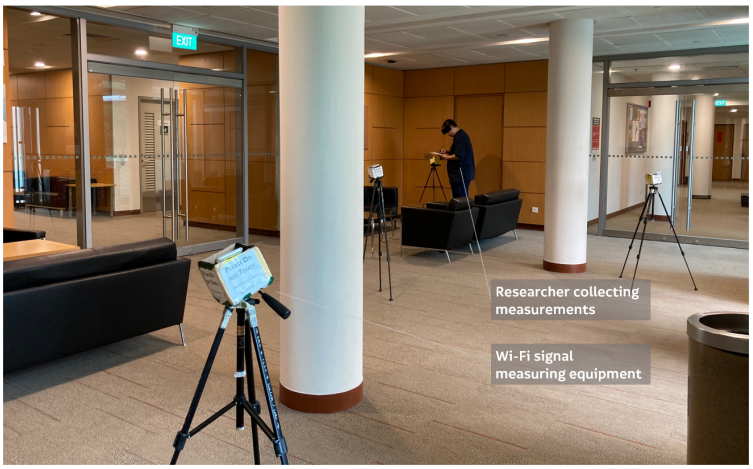

Fig. 13. An image captured while a research engineer collected Wi-Fi signal strengths to use as fingerprints for a Wi-Fi-based indoor localization system. The engineer used tripods (as shown) to house smart phones that were used to measure Wi-Fi signal strengths. For each data collection point, the Wi-Fi measurement required 1 minute. The floor map had 35 data collection points.

I USE OF SCALE-INVARIANT FEATURE TRANSFORM FOR IMAGE MATCHING

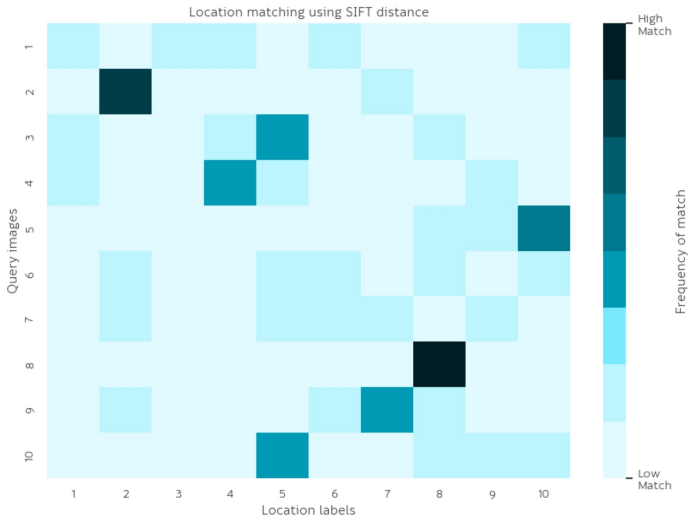

Fig. 14. Confusion matrix when attempting to use SIFT to match images to reference images captured at the locations.

We implemented SIFT to match images taken in study sites to the appropriate site based on reference images captured at the same site. To classify a query image, the algorithm iterated over every reference image at each study site. Using SIFT features descriptors, algorithm calculated a similarity score between the two images (query image and the reference image). The reference image with the highest similarity score was regarded as the best match and the study site label of the reference image was assigned to the query image. Figure 14 shows the confusion matrix of the assignment. Only sites 2 and 8 each had high accuracy of classification. Upon examination, both the sites had unique visual features. Site 2 had a large food outlet with black walls and signage. Site 8 had a long series of posters. We believe that based on the types and properties of entities present in the study site, it would not be reasonable to assume the presence of such distinguished features in all areas within a deployment site. Further, we observed that incorrect classifications had matched feature descriptors that were common to multiple study sites. Some such features were pillars, wall, ceiling and floor, and signage such as toilets. Within a large deployment site, we can expect uniformity in features of the built environment which can exacerbate the number of false mappings.

J INCORPORATING TIME-SERIES INFORMATION

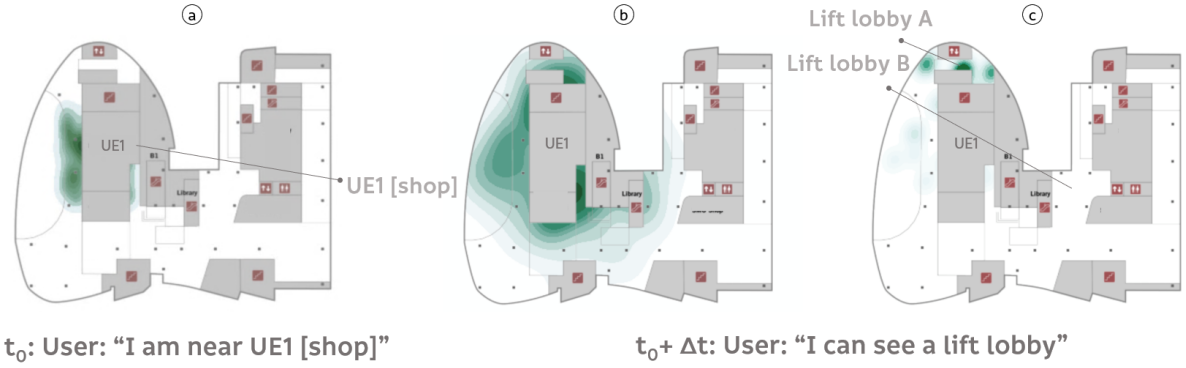

Fig. 15. Simulation of user motion starting from UE1 which is shop. (a): At time t_0, user mentions they are near UE1. We simulate the user's position based on this information. The kernel density plot of user's likely position is represented in (a). (b) & (c): After Δt seconds, at time = $t_0 + \Delta t$, user mentions they are near a lift lobby. Both lift lobby A and lift lobby B are potential candidates. Based on distance to the candidate lobbies, we assign weights to the user's likely position and adjust the distribution. Our simulation result indicates that lift lobby A is the more likely candidate as seen in (c).

In Fig.15.a, kernel density plot of particle distribution is depicted based on motion modeling of user who mentions entity UE1 at time t_0. In figures 15.b and 15.c, we incorporate this information about UE1 and the time duration Δt to generate new weighted distributions. We use the distance between the two candidate non-unique entities, i.e., lift lobby A and lift lobby B, and the weighted particles as a measure of likelihood of the non-unique entity being the correct candidate. Our simulation results show that "lift lobby A" is the more likely candidate, which is in accordance with the ground-truth used to model the simulation scenario.

K EXAMPLE INTERACTION ILLUSTRATING STRATEGIES USED BY CONVERSATIONAL AGENT

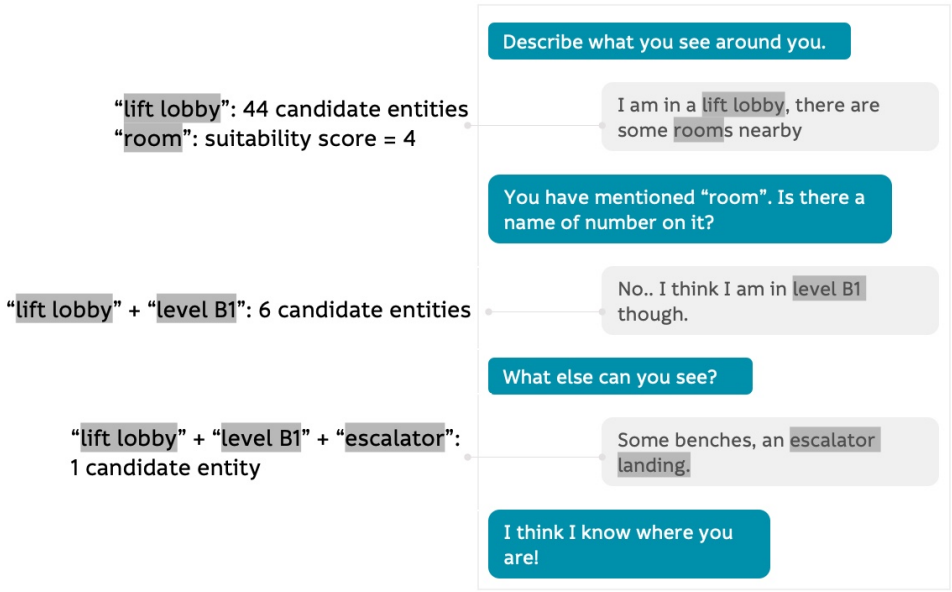

Fig. 16. A sample interaction with the conversational agent employing strategies of constraint programming and pursuing potentially valuable entities to arrive at an estimate.