The City as a Personal Assistant

Utku Günay Acer[†], **Marc Van den Broeck**[†], **Fahim Kawsar**[†*]
[†]Nokia Bell Labs, [*]TU Delft

ABSTRACT

Conversational agents are increasingly becoming digital partners in our everyday computational experiences. Although rich, and fresh in content, they are oblivious to users' locality beyond geospatial weather and traffic conditions. We introduce conversational agents that are hyper-local, embedded deeply into the urban infrastructure providing rich, purposeful, detail, and in some cases playful information relevant to a neighborhood. These agents are spatially constrained, and one can only interact with them once she is in close vicinity at street-level granularity. In other words, the city provides personal, stateful, spontaneous service to its citizens through the agents installed in urban landmarks. Drawing lessons from two user studies, we identify the requirements for this system. We then discuss the architecture of these agents that leverage covert communication channels and machine learning algorithms that run on the edge and wearable devices to offer meaningful conversational experience in urban settings.

CCS CONCEPTS

• **Human-centered computing** → **Sound-based input / output**; **Ubiquitous and mobile computing systems and tools**; *Mobile computing*.

KEYWORDS

conversational agents, location based services

ACM Reference Format:
Utku Günay Acer[†], Marc Van den Broeck[†], Fahim Kawsar[†*] . 2019. The City as a Personal Assistant. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2019 International Symposium on Wearable Computers (UbiComp/ISWC '19 Adjunct), September 9–13, 2019, London, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3341162.3349337

UbiComp/ISWC '19 Adjunct, September 9–13, 2019, London, United Kingdom

ACM ISBN 978-1-4503-6869-8/19/09…$15.00
https://doi.org/10.1145/3341162.3349337

1 INTRODUCTION

Personal assistants or conversational agents have become a significant part of computational experience in the last decade. These assistants understand spoken commands from the users to perform various tasks that are constrained by the applications installed on user devices. With the advancement in machine learning, the agents are able to understand the user speech in audio signals and convert text into speech. Once the user makes a vocal query, the agent sends raw audio signals to a remote server for natural language processing to recognize the intent of the user, and carries out the requested service using the API offered by the corresponding application. In addition to commercial-grade conversational agents (Alexa, Siri, Google, Cortana, etc.), specialized agents are used for accessing and interacting with digital services in many and diverse applications including customer experience [13], conversational commerce [14], medicine [15], and education [12].

Beyond weather and traffic conditions, the services provided by the agent do not consider the locality of users. Location based services, on the other hand, have also been widely investigated. Typically, once a user invokes a service request through a mobile device, the device acquires the current location and retrieves a service that is dependent on the user's spatial and temporal context from a remote server [5, 11].A large number of such applications have emerged offering a variety of experiences including navigation support, recommendation for venues, augmentation of a search for people, places, things, etc. These applications too often lack a locality view both temporally and spatially, i.e., information that is only available locally to local citizens. For example, consider a citizen would like to learn when the local events takes place such as the postman passing by, qualitative aspects on a neighborhood such as friendliness of the inhabitants, nearby shops that serve a particular product, the tourism related facts about a location that relate its historical and cultural background, etc. This extreme local information, today, unfortunately, is not available. Sensory and crowdsourcing systems are now being developed to accommodate such fine-grained view of an urban setting including both quantitative and qualitative fronts [1, 3, 6, 8].

We take a user-centered view and report two studies on the information affinity - different information citizens expect in an urban environment beyond what is available today and the modality of interaction for such information access.

Based on the results of these studies, we present a conversational agent system that offers local information to citizens. This system is embedded into the urban infrastructure. In addition, we move all the intelligence to the edge with machine learning algorithms either work in-situ or very close to the user. In other words, it is the city itself that provides the services to the users through this hyper-local conversational agent. This system offers spontaneous access to local information without relying on applications. We discuss the details of the main components of our system that relies on principles from covert communication, semi-stateful processing, and flexible conversation management.

2 CONTEXTUAL STUDIES

In this section, we report two studies. The first study aims at uncovering the variety of spatiotemporal information that a citizen wants to have from her neighbourhood. The objective of the second study is to understand this degree of information qualitatively, and extrapolating on the modality of the information accessibility.

Online Survey

The purpose of this survey is to understand the type of information users prefer to acquire concerning their neighbourhood in a spatiotemporal manner. We are mainly interested in determining a set of information types they are interested in and the modality of communication they prefer.

We have used Amazon Mechanical Turk[1] that allowed us to collect responses from 1992 participants in 5 continents (51.9% male, 48.1% female, age range 19-70 with 58.1% between 25 and 40). Within a range between 1 and 10 (1 lowest, 10 highest), 87% of the participants ranked their digital mindset above 5.

About their interest in local information, 90% of the participants indicate that they would like to be more informed about their surroundings in the city they live. While 81.4% said they can rather easily access local information through social media, search engines, billboards, city magazines, etc., 88.9% of the people said they would be interested in more effective approaches that do not require heavy filtering.

On *Interaction Modality*, we notice that most people mentioned that they use a conversational agent in public. It is the preferred medium for 24.3% of the participants. 40.5% of participants say that they would use it when they are alone whereas 13.9% remarks that they avoid using such an agent due to the noise in the city.

On *Information Affinity*, Figure 1 presents the types of information about their locality that the users would like to acquire. 45% of the users indicated they would make health-related queries to gather information including air quality, allergens, noise level, street cleanliness, etc. For 50.9%, community-related questions to provide information about new shops, social events, places with the best overall mood, crowded areas, etc. 59.2% of the people have said that they are like to inquire infrastructure related content such as public transportation, planned street works in an area. Safety related requests such as contacting the police, receiving and submitting warnings are important for 48% of the people. 54.3%, on the other hand, is interested in questions regarding the city including places to buy a certain product, places to go for certain activities, etc.

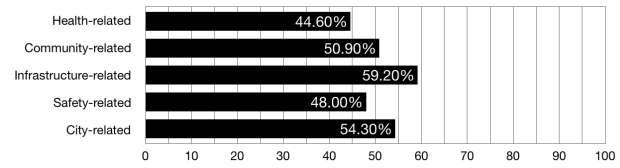

Figure 1: Types of Preferred Hyper-local queries

Semi-Structured Interviews

The survey offered us a quantitative view on information affinity, access frequency, and interaction modality. We wanted to understand the reasons behind the trends we reported earlier. To this end, we interviewed 21 individuals (13 men, 8 women, age range 25 - 72) following an interview technique called laddering[2] to uncover the core reasons behind users responses. We analysed the interview data by coding the individual responses using affinity diagram to derive final themes.

On *Information Affinity*, the participants essentially echoed our survey responses highlighting their desire for neighbourhood information concerning health outbreaks (e.g., pollen, flu, etc.), community events (e.g., street party, local school events, etc.), infrastructure related (e.g., construction schedule, noise management, etc.) and safety related (e.g., petty crime, etc.) Most participants have acknowledged that this system can help them engage with their neighbourhoods in a more informed manner. Some participants (n = 15) mentioned that they are not as aware of the city they live as they would have desired. They typically use social media or monthly city magazines to learn about events where they live in; however, filtering information generally is a problem.

On *Interaction Modality*, 8 participants have expressed that they prefer voice-based interaction for such information access. Multiple of them mentioned about conversational

[1] https://www.mturk.com/

[2] http://www.uxmatters.com/mt/archives/2009/07/laddering-a-research-interview-technique-for-uncovering-core-values.php

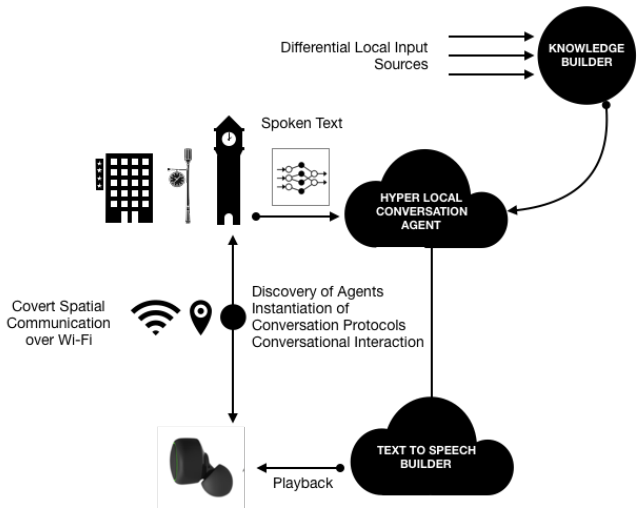

Figure 2: The architecture of the hyper-local conversational system.

agents commercially available today, e.g., Siri, Alexa, etc. as an example for interaction medium. Three of them have said they would have used the agent provided with their devices if it became more socially acceptable.

When asked about what they think of audio as a communication medium, sixteen people indicated that they like audio as a user interface if it is presented discreetly in a public setting, e.g., over a headphone.

Overall, all participants have expressed that they would have such a system. One concern people have is that it can be abused by a provider to flood users with advertisements or it can lead to information overload. As a remedy, some users noted that the system must not just push information, but it should only respond to specific queries. Another point of discussion was privacy. While people are willing to share limited data to receive, they are most interested in inquiries of public information such as availability in an enterprise, any event in the area, housing, health conditions, public transport, etc. Because the system provides spatiotemporal public information rather than a personal service that requires user data, it was perceived positively.

3 SYSTEM ARCHITECTURE

In this section, we provide a descriptive view of the proposed hyper-local conversational system. The architecture of this system is depicted in Figure 2 that is composed of a collection of agents. We envision the agents run embedded to local landmarks in an urban landscape such as a light post, a building, a tower, each responsible for its vicinity to provide spatio-temporally relevant information.

In this section, we discuss the details of the building blocks of the proposed system.

Interaction Unit

We use audio modality in user devices to receive any service. Users make requests via spoken commands. The requests propagate to the knowledge base through a WiFi AP over covert channels. Because we use covert WiFi communication channel instead of traditional protocols, we are not able to establish sessions to transport a large volume of data. Hence, we can send raw audio signal for analysis in a remote server. Instead, we deploy locally running machine learning algorithms to convert the user speech to text. Similarly, the respond to the user query also arrives in text. We use a speech builder to convert the text to speech and play it back to users. We use an off-the-shelf ear-worn device named eSense to capture the user queries and playback the responses [10]. A couple of eSense devices pair with a Raspberry Pi Zero W[3] through Bluetooth interface. The Raspberry Pi also provides a platform to run speech-to-text and text-to-speech engines. We use publicly available software packages such as Pocketsphinx [9] to speech to text conversation and tools like espeak and Pico Text-to-Speech for text to speech conversation.

Communication Unit

The agent instances have local connectivity with users through WiFi. They either run on WiFi Access Points (AP) that also has computation and storage capability or on other devices that are directly connected to APs. However, user devices need not associate with the AP to access the services provided by the agent. Instead, we propose to use covert channels to discover the agents and initiate conversations.

Recently, the extent to which WiFi management frames are utilized went beyond regulating the network operation. With 802.11u standard, management frames, a sub-category of management frames, facilitate *General Advertisement Service* (GAS) protocol that is used along with Access Network Query Protocol (ANQP) to roam around WiFi networks [7]. In addition, WiFi alliance has introduced WiFi-Direct technology where devices can create P2P links using management frames [4]. In this work, we adopt WiPush method to establish covert communication channels [2] that is based on management frames. While WiPush is designed to *push* notifications to user devices from the AP, on the other hand, in our system devices request service from the AP.

The WiFi devices use probe request frames to find out the available APs in their presence. The APs that hear the probe request to send a probe response that includes information about the AP. The capability to respond to hyper-local queries is included by the AP that hosts or connected to our conversational agent. Hence, through probe request/response frames, a user device is able to discover the agents.

[3]https://www.raspberrypi.org/products/raspberry-pi-zero-w/

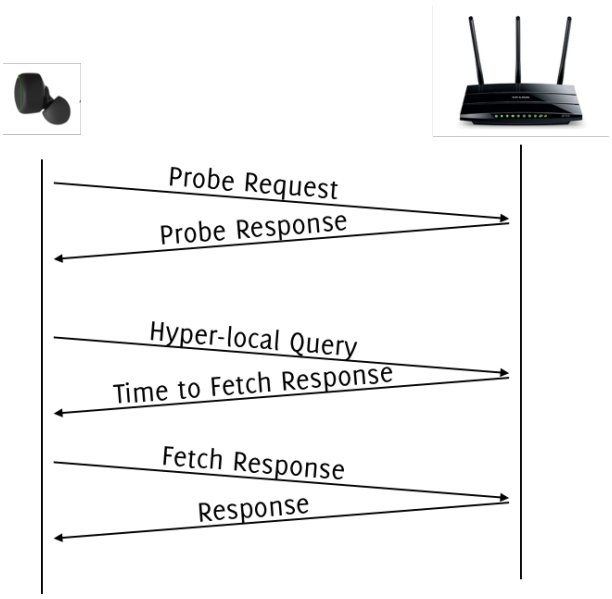

Figure 3: Protocol for Making Hyper-local Queries

The queries and the responses to them are exchanged, encapsulated in public action frames. Similar to WiPush, the action frames used in our system are able to carry a payload that is 2.3 KBytes.

When the user makes a query, the user device sends it to the agent nearby. Because it takes time to understand the query and prepare a response to the query, it can immediately respond to the user query, Also because the user devices are not associated with the AP, the AP needs the user devices to create the covert channel through a management frame that is followed by a message from the AP to the device. Hence, when the user makes a request, the AP will respond with a message that provides the expected duration after which the respond to the query is ready and the user can receive it. After that duration, the device sends another action frame that prompts the respond from the AP. This exchange of the messages is summarized in Figure 3.

To implement covert channels, we modify the *wpa_supplicant* daemon in user devices and in the *hostapd* daemon in APs, both working in the background and on the user space[4]. They handle the management of WiFi network, i.e. discovery, authentication, association and as such.

We use a Raspberry Pi 3 Model B+[5] to implement the AP side for the agent. The Raspberry Pi Zero W devices that the earpieces pair with have also a WiFi interface, and can be used to implement the user side of the covert channels.

[4]http://w1.fi

[5]https://www.raspberrypi.org/products/raspberry-pi-3-model-b-plus/

Knowledge Base

This is a question-and-answer component which enables a user to interact with the agent using a predefined dialogue base. Upon receiving the user request through the covert communication channel, the communication unit delivers the user request to the knowledge base.

The knowledge bases contextualizes the request with spatio-temporal information it has access to, i.e. beginning of the day in the city center where a number of people go to work vs a weekend afternoon in a city park. Taking such contextual information into account, the knowledge base runs natural-language-processing methods to understand the user intent and match it with a response that is populated with a set of situation- specific questions that the users can ask the agent.

Unlike similar conversation builders such as DialogFlow[6], this component runs in-situ on the agent. Therefore, the scope of the services and responds to users queries is limited to only their vicinity and users nearby. Within their vicinity; however, they have the complete knowledge.

In order to build the knowledge base, we plan use local datasets as well as crowdsourcing solutions. Crowdsourcing applications have already been widely deployed to collect fine-grained view of an urban setting including both quantitative and qualitative fronts. Proximity to WiFi APs have been used to retrieve the crowdsourcing tasks and upload responses to such tasks in [1]. Such methods can be beneficial to inquire and collect local information. In addition, we may use APIs from local resources to answer to dynamic queries.

Design Implications

Unlike other conversational agents, our conversation agents provide hyper-local service. The scope of the services and responds to users queries is limited to only their vicinity and users nearby. We consider remote requests are handled in servers through traditional methods hosted on the global Internet.

On the other hand, with an agent running on a large number of landmarks of the architectural landscape of an urban setting, we can present finely-grained spatiotemporal service to interested users.

Through such a wide coverage of agents, we strive to offer spontaneous and stateful interactions. As a user moves from one point to another that are served by different agents, the computation and data that is associated with serving the request move through these agents.

As an example, consider a user asks the agent to suggest a nearby restaurant with a particular item in its menu and with an available table. The agent checks the menu of every close-by establishment to see which ones serve the desired food and uses the APIs from the corresponding places to

[6]https://dialogflow.com/

inquire their availability. In another scenario, a person may want to know about the pollen count and CO_2 concentration at the very moment and location. When a user hears an ice cream truck, she may want to learn its location. A stateful query would be used in a tourism scenario where the city may use these agents to offer a guide experience to visitors or in an urban game such a scavenger hunt with digital assistance and/or tasks. In addition, the user requests may prompt action from public resources and local authorities such adjusting the brightness of the light for an evening gathering, or requesting the repair of a pothole. Such requests can be carried our without the need to connect to remote servers and exposing relevant APIs beyond the locality of the users, providing intrinsic geo-fences.

4 DISCUSSION AND OUTLOOK

In this paper, we present a work-in-progress conversational agent for urban settings. This hyper-local agent runs locally on urban landmarks and aggregates information from local sources to create a knowledge-base that contains every detail about its surroundings. The user devices opportunistically send queries to these agents leveraging the wide coverage of WiFi and retrieve the relevant local content using spoken commands. The service the users receive is spontaneous, flexible and stateful requiring user state to move across multiple agents reflecting the users' mobility.

We have performed two user studies to gauge potential users' interest for receiving local information. In either case, the participants have noted they would have used such a system to better engage with the urban setting they inhabit if some concerns are addressed. The current methods they use require various applications that depend on the content of the request and/or manual filtering of query results.

Such a system needs to address several issues. For example, a service from the agent should only be invoked by users to prevent unwanted advertisements. In addition, mechanisms that block information overload are necessary.

REFERENCES

[1] Utku Günay Acer, Marc van den Broeck, Claudio Forlivesi, Florian Heller, and Fahim Kawsar. 2019. Scaling Crowdsourcing with Mobile Workforce : A Case Study with Belgian Postal Service. In *Proceedings of the 2019 ACM International Joint Conference and 2019 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (UbiComp '19)*.

[2] Utku Günay Acer and Otto Waltari. 2017. WiPush: Opportunistic Notifications over WiFi Without Association. In *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous 2017)*. ACM, New York, NY, USA, 353–362. https://doi.org/10.1145/3144457.3144492

[3] Florian Alt, Alireza Sahami Shirazi, Albrecht Schmidt, Urs Kramer, and Zahid Nawaz. 2010. Location-based crowdsourcing: Extending crowdsourcing to the real world. *NordiCHI 2010: Extending Boundaries - Proceedings of the 6th Nordic Conference on Human-Computer Interaction*, 13–22. https://doi.org/10.1145/1868914.1868921

[4] Daniel Camps-Mur, Andres Garcia-Saavedra, and Pablo Serrano. 2013. Device-to-device communications with Wi-Fi Direct: overview and experimentation. *IEEE Wireless Communications* 20, 3 (June 2013), 96–104. https://doi.org/10.1109/MWC.2013.6549288

[5] Subhankar Dhar and Upkar Varshney. 2011. Challenges and Business Models for Mobile Location-based Services and Advertising. *Commun. ACM* 54, 5 (May 2011), 121–128. https://doi.org/10.1145/1941487.1941515

[6] Jakob Eriksson, Lewis Girod, Bret Hull, Ryan Newton, Samuel Madden, and Hari Balakrishnan. 2008. The Pothole Patrol: Using a Mobile Sensor Network for Road Surface Monitoring. In *Proceedings of the 6th International Conference on Mobile Systems, Applications, and Services (MobiSys '08)*. ACM, New York, NY, USA, 29–39. https://doi.org/10.1145/1378600.1378605

[7] Vishal Gupta and Mukesh Rohil. 2012. Enhancing Wi-Fi with IEEE 802.11u for Mobile Data Offloading. *International Journal of Mobile Network Communications & Telematics* 2 (08 2012). https://doi.org/10.5121/ijmnct.2012.2403

[8] Desislava Hristova, Afra Mashhadi, Giovanni Quattrone, and Licia Capra. 2012. Mapping Community Engagement with Urban Crowd-Sourcing. In *Proc. When the City Meets the Citizen Workshop (WCMCW)*. AAAI, Palo Alto, CA, USA, 14–19. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4749/5102

[9] David Huggins Daines, M Kumar, A Chan, A.W. Black, M Ravishankar, and Alexander Rudnicky. 2006. Pocketsphinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices, Vol. 1. https://doi.org/10.1109/ICASSP.2006.1659988

[10] Fahim Kawsar, Chulhong Min, Akhil Mathur, and Alessandro Montanari. 2018. Earables for Personal-Scale Behavior Analytics. *IEEE Pervasive Computing* 17 (07 2018), 83–89. https://doi.org/10.1109/MPRV.2018.03367740

[11] Axel Kupper. 2005. *Location-based Services: Fundamentals and Operation.* John Wiley & Sons, Inc., USA.

[12] Lindsay C. Page and Hunter Gehlbach. 2017. How an Artificially Intelligent Virtual Assistant Helps Students Navigate the Road to College. *AERA Open* 3, 4 (2017). https://doi.org/10.1177/2332858417749220 arXiv:https://doi.org/10.1177/2332858417749220

[13] Salvatore Parise, Patricia Guinan, and Ron Kafka. 2016. Solving the crisis of immediacy: How digital technology can transform the customer experience. *Business Horizons* 59 (05 2016). https://doi.org/10.1016/j.bushor.2016.03.004

[14] Nishant Piyush, Tanupriya Choudhury, and Praveen Kumar. 2016. Conversational commerce a new era of e-business. In *2016 International Conference System Modeling Advancement in Research Trends (SMART)*. 322–327. https://doi.org/10.1109/SYSMART.2016.7894543

[15] Myrthe Tielman, Mark Neerincx, Rafael Bidarra, Ben Kybartas, and Willem-Paul Brinkman. 2017. A Therapy System for Post-Traumatic Stress Disorder Using a Virtual Agent and Virtual Storytelling to Reconstruct Traumatic Memories. *Journal of Medical Systems* 41 (08 2017).