DuIVRS: A Telephonic Interactive Voice Response System for Large-Scale POI Attribute Acquisition at Baidu Maps

Jizhou Huang*
huangjizhou01@baidu.com
Baidu Inc.
Haidian District, Beijing, China

Haifeng Wang
wanghaifeng@baidu.com
Baidu Inc.
Haidian District, Beijing, China

Shaolei Wang
wangshaolei@baidu.com
Baidu Inc.
Haidian District, Beijing, China

ABSTRACT

The task of POI attribute acquisition, which aims at completing missing attributes (e.g., POI name, address, status, phone, and open/close time) for a point of interest (POI) or updating existing attribute values of a POI, plays an essential role in enabling users to entertain location-based services using commercial map applications, such as Baidu Maps. Existing solutions have adopted street views or web documents to acquire POI attributes, which have a major limitation in applying for large-scale production due to the labor-intensive and time-consuming nature of collecting data, error accumulation in processing textual/visual data in unstructured or free format, and necessitating post-processing steps with manual efforts. In this paper, we present our efforts and findings from a 3-year longitudinal study on designing and implementing DuIVRS, which is an alternative, fully automatic, and production-proven solution for large-scale POI attribute acquisition via completely machine-directed dialogues. Specifically, DuIVRS is designed to proactively acquire POI attributes via a telephonic interactive voice response system, whose tasks are to generate machine-initiative directed dialogues, make scripted telephone calls to businesses, and interact with people who answered the phone to achieve predefined goals through multi-turn dialogues. DuIVRS has already been deployed in production at Baidu Maps since December 2018, which greatly improves productivity and reduces production cost of POI attribute acquisition. As of December 31, 2021, DuIVRS has made 140 million calls and 42 million POI attribute updates within a 3-year period, which represents an approximately 3-year workload for a high-performance team of 1,000 call center workers. This demonstrates that DuIVRS is an industrial-grade and robust solution for cost-effective, large-scale acquisition of POI attributes.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**; • **Information systems** → **Location based services**.

*Corresponding author: Jizhou Huang.

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557131

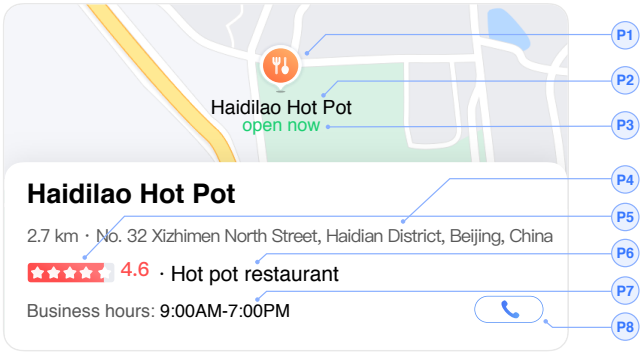


Figure 1: Example of POI information page at Baidu Maps.

KEYWORDS

POI attribute acquisition, interactive voice response system, task oriented dialogue system, knowledge acquisition, Baidu Maps

ACM Reference Format:
Jizhou Huang, Haifeng Wang, and Shaolei Wang. 2022. DuIVRS: A Telephonic Interactive Voice Response System for Large-Scale POI Attribute Acquisition at Baidu Maps. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3511808.3557131

1 INTRODUCTION

In commercial map applications, such as Baidu Maps, rich and exhaustive point of interest (POI) information (e.g., POI name, address, status, phone, open/close time, and other specifics) has always been a vital ingredient of ensuring user satisfaction with location-based services. Figure 1 shows an example of the POI multidimensional information page at Baidu Maps. As illustrated in this example, it provides users with an exhaustive overview of POI-centric information, which includes, for each POI, its coordinates and location (P1), name (P2), business status (P3), address (P4), ratings (P5), category (P6), open/close time (P7), and phone (P8). These POI attributes are of critical importance for the success of satisfying the needs of users, e.g., the attributes of status and open/close time are heavily exercised by users to make POI visiting decisions [26]. Therefore, in order to provide better experience and services, it is important to present users with as much exhaustive POI information as possible. One indispensable task in this context is POI attribute acquisition, which aims at completing missing attributes for a POI or updating existing attribute values of a POI.

The significance of this task is clearly evident from the facts that (1) new POIs emerge endlessly; (2) existing POIs are subject to change over time; and (3) POIs in certain sectors could be forced to

dramatically change during a public health emergency, such as the COVID-19 pandemic [13]. Further, recent statistics show that 74.5% of the POIs at Baidu Maps have been updated in 2020 [26], making it impractical to accomplish the attribute acquisition for hundreds of millions of POIs with manual efforts, which are labor-intensive and time-consuming. In order to drive productivity forward, several recent studies have attempted to develop new ways to acquire POI attributes by adopting street views [7] or web documents [26]. These solutions are typically framed in a three-stage paradigm: (1) extracting candidate attributes; (2) making links to a POI; and (3) sending the generated results to annotators for manual verification to ensure data quality. Although feasible, they have a major limitation in their applicability to large-scale production due to the labor-intensive and time-consuming nature of collecting data, error accumulation in processing textual/visual data in unstructured or free format, and necessitating a post-processing step with manual efforts. Moreover, despite their success in extracting salient POI attributes such as names and addresses, the practical applicability in applying them to deal with more attributes is typically limited because they are unable to generalize to POI attributes that are not conveyed by the collected images and documents.

Inspired by the idea of interactive voice response[1] (IVR), we suggest an alternative solution for large-scale POI attribute acquisition via a telephonic IVR system. An IVR system is a promising solution to automatically gather information from people via automated phone calls between humans and machines, where the dialogues are completely machine-initiated and the system is typically developed to detect predetermined keywords or special phrases uttered by people. The idea of IVR has been successfully implemented for information acquisition, such as Google Duplex [20], questionnaires [6], and knowledge acquisition [29].

However, it is highly challenging to build a telephonic IVR system for POI attribute acquisition that is characterized by robust stability, cost-effective performance, and industrial-grade reliability, mainly because of three facts. (1) A natural and coherent flow of human-machine conversations on the telephone heavily suffers from not only the performance of multi-turn dialogues with mixed-initiative interactions but also the errors introduced by automatic speech recognition (ASR) and natural language understanding (NLU). (2) The system is required to deliver low-latency voice responses, to people who answered the phone, with natural voices and speech intonations, while ensuring that the generated responses are reasonable and appropriate. And (3) the system is designed to have the capability of calling businesses and inquiring about diverse POI attributes at scale, which necessitates a fully automatic process while achieving an accuracy of human-level performance.

In this paper, we present our efforts and findings from a 3-year longitudinal study on designing and developing DuIVRS, which is an alternative, fully automatic, and production-proven solution for large-scale POI attribute acquisition via completely machine-directed dialogues. DuIVRS is designed to proactively acquire POI attributes via a telephonic IVR system, whose tasks are to generate machine-initiative directed dialogues, make scripted telephone calls to businesses, and interact with people to achieve predefined goals through multi-turn dialogues. Specifically, DuIVRS autonomously controls the dialogue flow by iteratively asking task-orientated questions w.r.t. a POI's attributes, and generating the next round of questions based on people's voice responses, until the goal is achieved or the phone is hung up by people.

DuIVRS has already been deployed in production at Baidu Maps since December 2018, which greatly improves productivity and reduces production cost of POI attribute acquisition. As of December 31, 2021, DuIVRS has made 140 million calls and 42 million POI attribute updates within a 3-year period, which represents an approximately 3-year workload for a high-performance team of 1,000 call center workers. This demonstrates that DuIVRS is an industrial-grade and robust solution for cost-effective, large-scale POI attribute acquisition.

Our main contributions to this problem are as follows:

- **Potential impact:** We suggest an industrial-grade and robust solution for cost-effective, large-scale POI attribute acquisition via a telephonic interactive voice response system. We document our efforts and findings from a 3-year longitudinal study on designing and developing DuIVRS, and we hope that it could be of potential interest to practitioners working with such problems. It is also hoped that this study could be a stepping stone to more viable IVR-based generalizations.
- **Novelty:** The design and development of DuIVRS are driven by the novel idea of calling businesses to inquire about POI information in a fully automatic manner without any manual intervention, so as to provide users with more accessible and timely POI information.
- **Technical quality:** DuIVRS has already been deployed in production at Baidu Maps for more than three years. The substantial real-world outcomes on productivity and efficiency demonstrate that DuIVRS is an industrial-grade and robust solution for large-scale POI attribute acquisition.
- **Scalability:** We examine the scalability of DuIVRS by showcasing its ability to adapt to rapidly support acquisition of new POI attributes.

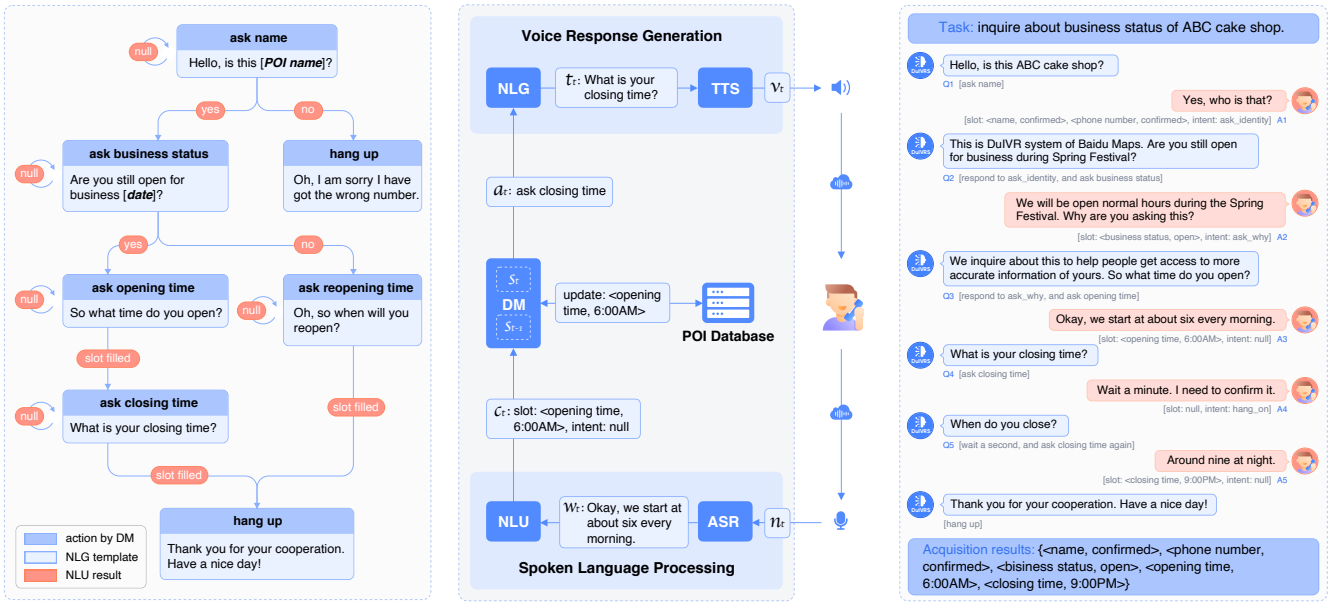
2 BACKGROUND AND RELATED WORK

Baidu Maps is one of the largest web mapping applications, which has a global POI database that provides information on about 180 million individual businesses and places. Rich and exhaustive POI attributes are the backbone of commercial map applications [26]. As such, they are required to exhibit high coverage and precision to better support visiting decision-making. In addition, rich and exhaustive POI attributes can also benefit the tasks of POI retrieval [8, 12, 14] and POI recommendation [4] at Baidu Maps. Next, we briefly review the closely related work in POI attribute acquisition from both academic exploration and industrial practice.

2.1 Academic Exploration

A related line of work has attempted to detect emerging and outdated POIs from web resources such as web snippets [5, 22] and tweets [31]. Additionally, some studies focus on detecting emerging POIs [23, 24] and updating POIs [7] from images. In spite of their capability of extracting salient POI attributes, such as name and location, they have two limitations for industry-level production.

[1] https://en.wikipedia.org/wiki/Interactive_voice_response



(a) Decision tree-based dialogue manager. **(b) System framework.** **(c) A real-world dialogue example.**

Figure 2: Illustration of DuIVRS.

First, they are unable to generalize to POI attributes that are not conveyed by the data sources. Second, a post-processing step with manual efforts is necessitated by the need to ensure data quality.

2.2 Industrial Practice

In Baidu Maps, four solutions that target POI attribute acquisition have been explored and deployed in production. The first solution (*S1*) utilizes unstructured text, such as web documents and online reviews, to extract POI names, status labels, and phone numbers, which has been presented in our previous work [26]. The second solution (*S2*) uses street views [7] to obtain POI names, coordinates, and addresses. Considering the limitations of post-processing and generalization, we further explore the third solution (*S3*) of manual verification. Before DuIVRS was deployed in production, we needed an initial and short-term solution as a stopgap, which would also be responsible for collecting training data. To this end, we built up a team of hundreds of call center workers, whose work was to manually call businesses to inquire about POI information. Practical results have shown that this solution is labor-intensive, costly, and more critically, hard to scale up. This is evident from the observations that: (1) the cost per call is about ¥1.5; (2) the maximum calls per day for each worker is only up to 200, due to the time pressure and concentration demands; (3) the highly repetitive and monotonous nature of the work is associated with an inability to maintain positive and productive conversations; and (4) the quality of acquisition results could vary widely across workers because it is difficult to standardize the work procedures through well-defined processes. Taking the above facts and constraints into consideration, we propose the fourth solution (*S4*) of DuIVRS to conduct large-scale POI attribute acquisition via a telephonic IVR system.

Taking the attribute of business hours as an example, we quantitatively compare the four solutions on the POIs in our POI database. Statistical results show that the POIs that have been updated with the correct attribute values of business hours by *S1*, *S2*, *S3*, and *S4* (DuIVRS) account for 4.6%, 0.9%, 3.7%, and 9.7%, respectively. This shows that DuIVRS is a production-proven solution for large-scale POI attribute acquisition.

3 DuIVRS

We aim to achieve five main design objectives in building DuIVRS. (1) The system is able to work in a fully automatic manner without any manual intervention while achieving human-level performance on POI attribute acquisition, which necessitates autonomously calling businesses to inquire about POI information, extracting attribute values from utterances, and updating the verified POIs. (2) The system is required to deliver coherent, natural, and low-latency voice responses with reasonable accuracy. (3) The system is able to inquire about information that is necessarily absent from publicly available data sources. (4) The acquisition processes can be standardized, such as generating inquiries and replies from within a pre-written script template. (5) The system can deal with large-scale production and easily be scaled-up with a low cost. Motivated by those objectives, we adopt a modular framework that decomposes the whole system into five cascaded components, which benefits from the recent advances in IVR systems [1, 18, 25, 33] and task-oriented dialogue systems [3, 10, 11, 16, 19, 28, 32].

3.1 Overview

Figure 2 depicts the overall framework of DuIVRS, which is accompanied by a real-world example. In this example, DuIVRS proactively makes a telephone call to a business to inquire about POI information. For each round t of the dialogue, DuIVRS conducts the five components in sequence: (1) the automatic speech recognition (ASR) component, which transcribes the speech of people into

text u_t; (2) the natural language understanding (NLU) component, which parses u_t into a machine-readable, structured semantic representation s_t including user intents and slot values; (3) the dialogue management (DM) component, which interprets s_t and decides the next action a_t to take; (4) the natural language generation (NLG) component, which transforms a_t into a text response r_t; and (5) the text to speech (TTS) component, which synthesizes natural-sounding speech o_t from r_t. DuIVRS integrates multiple rounds of dialogue to complete the whole POI attribute acquisition process without any manual intervention.

To address the three challenges (see §1 for details) and achieve the five main design objectives (§3), we take recent advances in task-oriented dialogue systems and apply a step-wise optimization approach to incrementally achieve performance improvement. Specifically, we perform a component-by-component ablation study and refine them individually, including (1) fine-tuning the ASR component with domain-specific data to adapt to a call center scenario (§3.2); (2) incorporating Chinese pinyin into the representations of utterances, to alleviate the impact of ASR errors and improve intent detection accuracy (§3.3); (3) embedding human expert knowledge into a rule-based finite-state DM component, to assure the ability to conduct fluent and coherent conversations (§3.4); (4) taking into account the speaking style of call center workers in designing the NLG component, to make the generated spoken-style text more natural and reasonable (§3.5); (5) generating closed-ended questions to reduce variability in user responses and narrow down the answers, which enables us to obtain promising intent detection and slot filling results (§3.5); (6) identifying the multiple user intents and generating rational responses to support mixed-initiative interactions, which makes the conversation as natural and comfortable as possible to humans, and hence it flows smoothly (§3.3-§3.5); (7) fine-tuning the TTS component with a pronunciation dictionary consisting of geographic entities (e.g., POI name and address), to synthesize natural-sounding speech (§3.6); and (8) applying a set of deployment strategies, such as storing high-frequency sentences in a shared cache, to facilitate large-scale, low-latency services (§3.7).

3.2 Automatic Speech Recognition

Problem Statement. ASR aims at automatically transcribing the speech signals into text. Compared to general-purpose applications, the development of ASR for DuIVRS is challenged by four problems. (1) *Chinese accents.* Many Chinese people speak non-standard varieties of Mandarin with regional accents ("Difang Putonghua" in Chinese), such as Shanghai, Sichuan, Guandong, and Fujian, which inevitably results in possible mispronunciations and accented speech. (2) *Background noise.* The possible noise sources (e.g., talk, music, and television) in business environments (e.g., restaurants and hotels) generate the acoustic ambient noise during calls. (3) *Sound quality.* The sound quality of output signals varies greatly among handheld devices. (4) *Low latency.* The low latency requirement of real-time speech interaction necessitates transcribing the speech signals as rapidly as possible.

Our Practice. Although the generic ASR systems built at Baidu have achieved significant improvements in accuracy and latency, the recognition performance drops dramatically when dealing with conversational telephone speech due to the combination of three challenges (i.e., Chinese accents, background noise, and sound quality) and the inability to correctly transcribe domain-specific words and phrases, such as POI names. To address this problem, the speech technology team fine-tuned a domain-specific ASR model with over 1,000 hours of call center data. In addition, they further integrated a geo-specific language model trained on our complete POI database into the fine-tuned ASR model to improve recognition accuracy of domain-specific words and phrases.

3.3 Natural Language Understanding

Problem Statement. NLU aims at identifying user intents and extracting semantic slots, which mainly consists of two subtasks: intent detection and slot filling. For example, given an utterance, "We will be open normal hours during the Spring Festival. Why are you asking this?", the output of NLU consists of a user intent "*ask_why*" and a slot "*<business status, open>*". The development of NLU for DuIVRS is challenged by three major problems. (1) *ASR errors.* Even fine-tuned ASR cannot escape from errors, which will be propagated and inevitably have effects downstream. (2) *Unpredictable interactions.* The responses to inquiries about POI information may not be in line with what is anticipated. For example, people may respond with natural language sentences rather than keywords. Sometimes, they may raise their questions rather than answer ours. (3) *Performance bottleneck.* The failure of NLU leads to inappropriate responses of DuIVRS and causes the dialogue to break down. However, it also remains a bottleneck in correctly interpreting the meaning of an utterance, due to noise, newly emerging intents, and imbalanced classification categories.

Our Practice. To successfully sustain the conversation flow over time and enable smooth turn-taking interactions during the call, it is important to accurately understand the user intents and generate appropriate responses. We address the above-mentioned three challenges through a series of refinements, including: (1) asking questions in a task-friendly manner to narrow down the answers into a more manageable set; (2) developing a multi-label classification network to identify multiple intents from user utterances and further incorporating Chinese pinyin into it to mitigate the impact of ASR errors; and (3) building effective feedback loops to iteratively collect failed question-answer pairs, which are used to uncover new intents and augment the training data.

First, to obtain promising intent detection and slot filling results, we constrain DuIVRS to ask closed-ended rather than open-ended questions. In addition, we decompose the task of inquiring about a composite attribute of a POI into a set of simple questions that elicit "yes-or-no" answers or brief answers (see §3.5 for details). For example, when inquiring about the slot value of business hours, we decompose it into two short questions by first asking for opening hour and then asking for closing hour. As a result, the answers can be significantly narrowed down, which enables us to obtain a task-friendly input for NLU. The effectiveness of this method is also confirmed in [21], which shows that asking for constraining questions can obtain better NLU performance than open-ended questions during a human-machine dialog.

Second, we regard the detection of user intents as a multi-label classification task, where the key challenge is to mitigate the impact of ASR errors. To address this challenge, a widely-used approach

advocates the development of an additional error correction component in between ASR and NLU to deal with the possible errors in the transcribed text. However, the biggest problem in dealing with ASR errors in developing DuIVRS is that the majority of ASR errors are phrase-level errors, which may significantly change the meaning of the transcribed text. As a consequence, it is impractical to label large-scale training data, because the transcribed text with phrase-level errors is often incomprehensible to human annotators. This makes it difficult to develop an additional error correction component that exhibits promising performance. In our practice, we use an alternative and implicit way to tackle this problem.

Our statistical analysis shows that the largest proportion of ASR errors involved homophone words or words with similar pronunciation, which is also a special phenomenon in Chinese ASR systems. For example, even though the two Chinese words "北京" and "背景" share the sample Chinese pinyin "bei jing", they have completely different meanings. Moreover, some people speak different dialects of Chinese (e.g., they may have trouble with "n" and "l", and thus cannot differentiate "nan" and "lan"), which inevitably leads to mis-transcribed words. Based on these observations, we develop a dual fusion attention-based convolutional network (DFAC) to detect user intents, which makes use of the pronunciation information of Chinese words to help recover from the ASR errors.

Figure 3 shows the network structure of DFAC. Given an utterance U_t consisting of a sequence of characters $w_1 w_2 \cdots w_n$, we transform them into character embeddings $e = [e_1, e_2, \cdots, e_n] \in \mathcal{R}^{d \times n}$ and Chinese pinyin embeddings $p = [p_1, p_2, \cdots, p_n] \in \mathcal{R}^{d \times n}$. e and p are then sent to the CNN networks to obtain the text representation $z_e \in \mathcal{R}^m$ and pinyin representation $z_p \in \mathcal{R}^m$, respectively. To explicitly model relation and interaction between e and p, we further introduce a fusion attention mechanism that is defined as:

$$\mathbf{I} = e \oplus p \oplus (e - p) \oplus (e * p), \tag{1}$$

$$\mathbf{z} = \mathbf{W}_u \tanh(\mathbf{W}_v \mathbf{I}), \tag{2}$$

$$\alpha_i = \frac{exp(\mathbf{z}_i)}{\sum_{j=1}^{n} exp(\mathbf{z}_j)}, \tag{3}$$

$$\mathbf{o} = \sum_{i=1}^{d} \sum_{k=1}^{n} (\alpha_k \cdot e_{ik} + (1 - \alpha_k) \cdot p_{ik}), \tag{4}$$

where $\oplus$ is the concatenation operator which results in $\mathbf{I} \in \mathbb{R}^{4d \times n}$, $\alpha \in \mathbb{R}^{n \times 1}$ is the attention weight, and $\mathbf{o} \in \mathbb{R}^d$ is the aggregated representation of e and p.

z_e, z_p, o are concatenated into a vector $\mathbf{H} = [z_e, z_p, o] \in \mathbb{R}^{2m+d}$. $\mathbf{H}$ is then fed into a sigmoid layer to make the final prediction by $\hat{y} = \sigma(\mathbf{WH})$, where $\mathbf{W} \in \mathbb{R}^{|C| \times (2m+d)}$ and $|C|$ is the total number of intents. We use $y \in \mathbb{R}^{|C|}$ to denote the ground-truth intents of an utterance, where $y_i \in \{0, 1\}$ indicates whether intent i appears in the utterance or not. DFAC is trained with the multi-label cross entropy loss:

$$\mathcal{L} = \sum_{c=1}^{|C|} y^c \log(\hat{y}^c) + (1 - y^c) \log(1 - \hat{y}^c). \tag{5}$$

Third, the DFAC model trained with manually annotated data can hardly deal with newly emerging intents that have not been encountered before. Therefore, the performance of DFAC inevitably

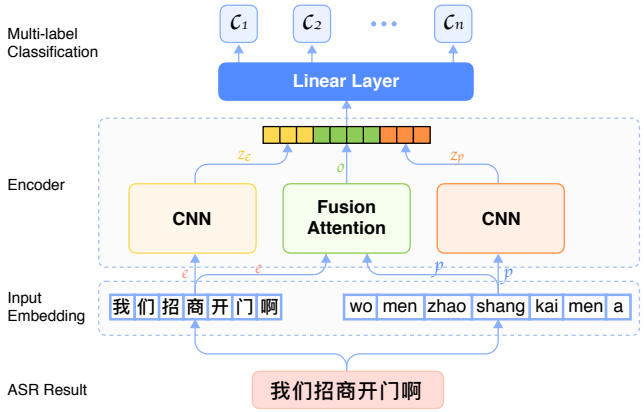


Figure 3: Illustration of DFAC.

decreases over time. To address this issue, we build effective feedback loops to iteratively improve its performance. Specifically, we accumulate the failed question-answer pairs of DuIVRS, and routinely cluster these answers into different intents based on the assumption that utterances with the same intent share more similar context features, such as keywords. Next, we manually analyze the clustering results to discover new intents, annotate sufficient data, augment the training data, and re-train DFAC to achieve superior performance. For other components of DuIVRS, we also apply this idea to routinely improve their performance.

The call center dialogue logs are used to train DFAC as follows: (1) We randomly sample and annotate 20,000 answers to initialize DFAC; (2) we iteratively augment the training data; and (3) we re-train DFAC to achieve superior performance. When adapting to new attribute acquisition tasks, the above three steps are applied to continuously optimize its performance.

The success of deploying DuIVRS in a fully automatic manner without any manual intervention necessitates filling the slots as accurately as possible. Likewise, the design goal of DuIVRS is to accomplish POI attribute acquisition tasks in a fully automatic manner, which also necessitates filling the slots as accurately as possible. To this end, the NLG component is constrained to generate closed-ended questions in a one-fact-per-question fashion. In this way, people only need to provide "yes-or-no" answers or brief answers. As a consequence, slot values can be easily inferred from the user-provided answers, which allows us to adopt a widely-used template-based method [2] to extract slot values. Specifically, we manually create high-frequency templates, and the feedback loop strategy is applied to iteratively cluster and generate new templates after launching. Our practical results demonstrate that this solution can achieve promising performance on the slot filling task.

3.4 Dialogue Management

Problem Statement. DM aims at keeping track of the overall interaction with a view to ensure steady progress towards task completion. Specifically, it is responsible for maintaining the interaction state, deciding the next action of "what to say", and clarifying potential misinterpretation of answers. For example, given a user intent "*ask_why*" and a completed slot "*<business status, open>*" extracted by NLU, the next action DuIVRS should take is to "respond *ask_why*" and then "ask opening time". The development of DM for DuIVRS

is challenged by two major problems. (1) *Reliability*. To ensure successful task completion and obtain information as much as possible, it is important to accurately keep track of the overall interaction, which plays a vital role in enabling DuIVRS to deliver coherent and natural responses. (2) *Robustness*. Errors inevitably exist in understanding what people say. Moreover, user responses are highly diversified. For example, some people prefer providing brief and straightforward answers, while some raise their questions before answering ours. Therefore, DM should explicitly handle all possible unusual responses that may occur during calls.

Our Practice. We address these challenges through a series of refinements, including: (1) decomposing an acquisition task into a succession of simple subtasks to obtain a better yield; (2) developing a constrained and lightweight decision tree-based model to accurately shepherd the interaction towards task completion, which is conducive to the standardization of the acquisition process; and (3) developing a mixed-initiative dialogue strategy to explicitly handle possible unusual responses.

First, an acquisition task is decomposed into a succession of simple subtasks, which is beneficial to obtain task-friendly user responses for NLU and control the acquisition process. This refinement is inspired by the production process of high-performance call center workers who prefer to decompose a task into a succession of simple subtasks to improve productivity and efficiency. For example, when performing a task of inquiring about business information, they decompose it into four subtasks, including: ask name, ask business status, ask opening time, and ask closing time.

Second, to fully standardize the acquisition process and facilitate automated information gathering, we develop a constrained and lightweight decision tree-based model for task decomposition and organization, which is designed to decompose an acquisition task into subtasks and structure subtask dependencies. Specifically, we employ a hierarchical approach to model the dependencies between different subtasks, and convert them into a decision tree structure by mapping each subtask to a specific node. This approach enables DuIVRS to facilitate task completion by concurrently monitoring goal progress on the task at hand and providing appropriate guidance. Figure 2a showcases the decision tree structure of a task that aims at inquiring about business status. For example, if a slot value is successfully obtained, the current node will be marked as completed, and meanwhile, the next action is decided by using a decision tree with the completed slots and values. Otherwise, it would repeat the current action once, if nothing has been obtained. The dialogue is finished if the leaf node is executed or people hang up the phone.

Third, user responses are highly diversified, and unusual responses inevitably occur during calls. Therefore, DM should explicitly handle all possible responses, such as questions raised by people. To this end, we expand DuIVRS to support a mixed-initiative interaction by detecting, interpreting, and responding to the user intents that have been engaged in historical dialogues between people and our call center workers. To make the decision trees more readable and easily renovated by non-specialists, we decouple the generation of responses to user-initiative questions from the decision tree-based dialogue manager. In our practice, the detected user intent by NLU and the generated action by DM are concurrently passed to the NLG component to produce a response that consists of two primary parts: (1) a natural language sentence that addresses the question requested by the user, and (2) a machine-initiative question that directs the next action or repeats the current action if the accompanied question was not answered.

Figure 2c shows a real-world dialogue example conducted by DuIVRS. We can see that DuIVRS has successfully completed the acquisition task through making scripted calls, handling unusual user responses (see A1 and A2 in Figure 2c for examples), and interacting with people through multi-turn dialogues.

3.5 Natural Language Generation

Problem Statement. NLG aims at converting the action produced by the DM component into a natural language response, which takes as input "what to say" and determines "how to say". For example, the action "*ask opening time*" can be uttered as "So what time do you open?". The ability of DuIVRS to generate reasonable and appropriate responses can lead to better user engagement. To achieve this goal, the NLG component needs to be capable of delivering the following advantages. (1) *Controllability*. To motivate users to provide more substantive and actionable information through slot values, it is important for NLG to generate responses that are grammatically fluent, unambiguous, and easy to understand for people who answered the phone. (2) *Task-friendly questions*. We observe that, based on the user behavior of call center dialogues, if a question is vague, contemplative, or ambiguous, it would inevitably lead to long responses, multiple slot values, and lower user engagement, making it difficult for NLU to effectively handle them. Motivated by this, it is important to ask questions that are friendly to people and downstream tasks, which enables us to better identify user intents and obtain promising slot values. (3) *Spoken-style questions*. Spoken language and written language are different in cognitive and linguistic properties. To prompt users to provide information in a user-friendly manner, it is important to generate spoken-style questions rather than written-style questions.

Our Practice. We address these challenges through a series of refinements, including: (1) developing a template-based approach to generate responses; (2) constructing closed-ended question templates that impose the constraints of simplicity and brevity; (3) generating questions in a one-fact-per-question fashion; (4) deriving spoken-style questions that return high-quality, user-provided answers from call center dialogues; and (5) imitating typical abbreviations for lengthy expressions or frequently used phrases.

First, a template-based approach is developed to generate questions and responses. The templates are constructed considering constraints of simplicity and brevity, which permit generating closed-ended questions (e.g., "Hello, is this [*POI name*]?") or one-fact-based questions (e.g., a question of asking business hours is decomposed into two one-fact-based questions, "What time do you open?" and "What is your closing time?") that elicit "yes-or-no" answers or brief answers (e.g., "six"). In this way, DuIVRS is able to generate simple and concise questions, which are more friendly to users as they can think without effort and respond quickly. Consequently, the variability in user responses can be drastically reduced, and the answers can be significantly narrowed down. In addition, if user intents (e.g., *ask_why*) are detected during calls, we first get a

pre-written sentence that addresses such user concerns, and then we put it before the generated machine-initiative question as the final response (see Q2 and Q3 in Figure 2c for examples).

Second, we use the call center dialogue logs to mine "the wisdom of the crowds" for question template learning, which enables us to extract questions that have been proven to be effective for obtaining slot values from users. Specifically, we derive spoken-style questions that achieve high success rate of user-provided answers from the dialogue logs, and then we learn templates from these questions using the widely-used rule-based approach [27] due to its robustness and ability to produce high-quality results in practical applications. In this way, the response diversity can also be increased because multiple templates are built for individual actions.

Third, the POI attributes, such as name and address, are mostly in written style rather than spoken style, which would not necessarily apply to naturally spoken utterances. Moreover, inserting written-style phrases into a template would inevitably decrease the fluency and naturalness of spoken utterances [17]. Therefore, to generate spoken-style questions, it is critical to imitate typical abbreviations for lengthy expressions or frequently used phrases. To address this issue, we manually construct a set of written-to-spoken style conversion rules by learning from large amounts of questions among call center workers. For example, the written-style POI name "*Meizhou snack (Wucaicheng Shopping Center store)*" is simplified into "Meizhou snack in Wucaicheng".

3.6 Text To Speech

Problem Statement. TTS aims at synthesizing natural-sounding speech from the generated text of NLG. Compared to general-purpose applications, the development of TTS for DuIVRS is challenged by three problems. (1) *Human-sounding speech.* It has been shown that a machine-sounding voice is typically perceived as unnatural and unpleasant [9]. Therefore, to improve user engagement during calls, it is important to generate human-sounding speech. (2) *Geographic entity pronunciations.* The lack of typical letter-to-sound mapping rules and text-to-phoneme transformations poses a major challenge to generate the pronunciations for a large percentage of geographic entities (e.g., POI name and address). (3) *Low latency.* The low latency requirement of real-time speech interaction necessitates synthesizing the speech as rapidly as possible.

Our Practice. Although the generic TTS systems built at Baidu have achieved significant improvements in naturalness and latency, the synthesis quality drops dramatically when dealing with geographic references within text due to the inability to synthesize speech with acceptable quality for diverse geographic entities and other domain-specific words. To address this problem, the speech technology team built a specific pronunciation dictionary for geographic entities, and it further trained a domain-specific TTS model with over 10,000 paired speech and text collected from call center data.

3.7 Deployment Strategies

In addition to model optimization and component refinement, we also explored practical deployment strategies to meet the large-scale production efficiencies and capabilities of industry. Here we mainly introduce those for low latency processing and scalability.

Closed-ended questions lead to high-frequency answers (e.g, "Yes" and "Of course"). Based on this observation, we store high-frequency answers and their contextual question templates in a shared cache. During the call, DuIVRS directly extracts the subsequent question template from the shared cache without calling the NLU and NLG components when the high-frequency answer is triggered. In addition, a question template generally consists of fixed phrases and non-fixed chunks (e.g., "Hello, is this [*POI name*]?"). Based on this fact, we use asynchronous synthesis mechanism to further reduce the delay. Specifically, the voice segments of fixed phrases are stored in a shared cache in advance. During the call, once a response is generated by NLG, DuIVRS starts to play the pre-recorded voice segments of preceded phrases in the cache, and meanwhile calls TTS to synthesize the non-fixed chunks. Our statistics show that the mean latency of DuIVRS decreased from 23 milliseconds to 1 millisecond by applying the caching strategy.

We build a dialogue customization and management platform to achieve scalability and flexibility. It integrates each component of DuIVRS into a visual interface for presenting and managing user intents, templates, and decision trees. This enables us to focus on the design of dialogue structure and process itself, which in turn significantly simplifies the development and implementation of new attribute acquisition tasks. Our statistics show that the mean development period of model optimization and component refinement decreased from 4 days to 1 day after applying this platform.

4 EXPERIMENTS

4.1 Evaluation Metric

As detailed in §3.1, the DuIVRS system is cascaded over multiple components. As a consequence, the overall performance of DuIVRS is limited by the component with the worst performance. To this end, we use an end-to-end, system-level evaluation method to address the challenge in obtaining a list of locally optimal components rather than yielding a globally optimal system. Specifically, we use task success rate (*TSR*) to evaluate the effectiveness of DuIVRS for POI attribute acquisition, which is calculated by: $TSR = \frac{N_C}{N_T} \times 100\%$, where N_T is the total number of questions asked by DuIVRS, and N_C is the total number of slots that are successfully extracted. For example, DuIVRS asked 5 questions in the dialogues presented in Figure 2c. Among which, 4 questions were answered and 4 slots were successfully extracted. The *TSR* of this dialogue example is $4/5 = 80\%$. *TSR* is an overall evaluation criterion that reflects our business goal (i.e., cost reduction through improved productivity), as well as it drives optimization that incorporates accuracy, robustness, and efficiency.

4.2 Production Gains of DuIVRS

DuIVRS has already been deployed in production at Baidu Maps since December 2018, which works in a fully automatic manner without any manual intervention while achieving human-level performance on POI attribute acquisition. As of December 31, 2021, DuIVRS has made 140 million calls and 42 million POI attribute updates within a 3-year period, which represents an approximately 3-year workload for a high-performance team of 1,000 call center workers. This shows that DuIVRS can greatly improve productivity and reduce production cost of POI attribute acquisition.

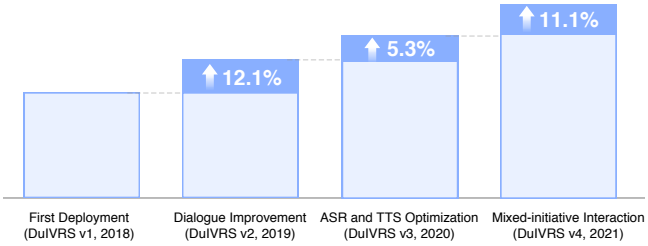


Figure 4: *TSR* gains of improved versions of DuIVRS.

In addition, the acquisition process of DuIVRS can be triggered periodically (e.g., at daily, weekly or monthly time intervals) or on demand (e.g., after the creation of POIs in the database), which makes it powerful in handling emergency situations, such as the COVID-19 pandemic, during which POIs could be forced to dramatically change [13, 30]. For example, many businesses were closed or reopened regarding COVID-19 and the approved prevention measures. To help people get timely access to business changes with Baidu Maps, we employ DuIVRS to rapidly acquire up-to-date POI information regarding the daily necessities of life, and 3.6 million updates were successfully applied during the first quarter of 2020. These updates helped users suffer as little inconvenience as possible when making visiting decisions during the COVID-19 pandemic.

4.3 System Evolution

Building an industrial telephonic IVR system for POI attribute acquisition that is characterized by robust stability, cost-effective performance, and industrial-grade reliability is not something that can be done overnight; it is the culmination of step-wise processes and a series of consecutive refinements. DuIVRS has been updated three times since the initial system was deployed online in December 2018. For each time, before being launched in production, we would conduct an online A/B testing. Specifically, we would deploy the new version online and make it randomly serve 5% of the call tasks. During the period of A/B testing, we would monitor the performance of the new version and compare it with the performance of former version that has been deployed successfully online. This period conventionally lasts for at least one week. Figure 4 shows the performance gains of three versions of DuIVRS.

First Deployment (DuIVRS v1). The initial version of DuIVRS is deployed online in December 2018, which consists of five components: (1) the generic ASR system built at Baidu; (2) the NLU component that only deals with task-related intents; (3) the constrained and lightweight decision tree-based DM that is able to decompose an acquisition task into a succession of simple subtasks; (4) the template-based NLG component; and (5) the generic TTS system built at Baidu. *DuIVRS v1* validated that it is feasible to develop a telephonic IVR system for large-scale POI attribute acquisition.

Dialogue Improvement (DuIVRS v2). In 2019, we deployed *DuIVRS v2* that mainly optimized the dialogue strategies through a series of refinements, including: (1) constructing closed-ended question templates that are friendly to people and downstream tasks; (2) generating questions in a one-fact-per-question fashion; (3) deriving high-quality, spoken-style questions from call center dialogues; and (4) imitating typical abbreviations for lengthy expressions or frequently used phrases. The online A/B testing result shows that *DuIVRS v2* obtains significant (absolute) improvement by 12.1% *TSR* compared with *DuIVRS v1*.

ASR and TTS Optimization (DuIVRS v3). The generic ASR and TTS systems built at Baidu are unable to deal with most of the geo-specific words and phrases. In 2020, we deployed *DuIVRS v3* that mainly optimized the ASR and TTS components through a series of refinements, including: (1) fine-tuning a domain-specific ASR model, which improved the keyword spotting accuracy from 71.6% to 84.5%, and (2) fine-tuning a domain-specific TTS model, which significantly improved the naturalness of synthesized speech. The online A/B testing result shows that *DuIVRS v3* obtains significant (absolute) improvement by 5.3% *TSR* compared with *DuIVRS v2*. This suggests that it is necessary to develop domain-specific ASR and TTS models for building a task-oriented telephonic IVR system.

Mixed-initiative Interaction (DuIVRS v4). We observed that unusual responses sometimes occur during calls. For example, people may raise their questions rather than answer ours. If there was no response to their questions or the response was not accepted by them, then the motivation of people to continuously interact with DuIVRS can be drastically weakened. To better maintain the motivation to interact with DuIVRS, it is important to support mixed-initiative interactions. In 2021, we deployed *DuIVRS v4* that supported mixed-initiative interactions through a series of refinements, including: (1) developing a multi-label classification network to identify user intents from utterances, and further, incorporating Chinese pinyin into it to mitigate the impact of ASR errors, and (2) building an effective feedback loop that is designed to iteratively collect failed question-answer pairs, uncover new intents, and improve dialogue success rates. *DuIVRS v4* is able to handle 16 frequent user intents and achieves a micro-F1 of 97.9% for multi-label intent detection. After this refinement, *DuIVRS v4* can successfully handle over 96% of utterances that cannot be addressed by *DuIVRS v3*. The online A/B testing result shows that *DuIVRS v4* obtains significant (absolute) improvement by 11.1% *TSR* compared with *DuIVRS v3*.

4.4 Analysis

Robustness. We evaluate the robustness of DuIVRS in terms of credibility and accuracy of the acquisition results.

To investigate whether people may provide incorrect information, we deliberately inject incorrect information (e.g., wrong POI name or address) into the generated questions of NLG, then perform calls with DuIVRS, and analyze their answers. We manually evaluate 500 question-answer pairs of this test. Statistics show that 64% of answers are determinate "No", and the rest are confirmation responses (e.g., "Could you repeat that, please?"), while no one provides incorrect slot values. This shows that the acquisition results are highly reliable, which confirms the feasibility of using DuIVRS for large-scale POI attribute acquisition.

To evaluate the accuracy of acquisition results, we ask call center workers to manually verify the extracted slot values by referring to the voice clips of 1,000 randomly sampled question-answer pairs. Statistical results show that 99.2% of extracted slot values are consistent with those provided by users, which achieves human-level performance. This enables us to deploy DuIVRS in a fully automatic manner without any manual intervention.

Table 1: Comparison between human and DuIVRS.

	Human	DuIVRS
TSR	89.6%	79.9%
cost per call	¥1.5	< ¥0.2
calls per day	≤ 200	no limitations
standardization	×	√
stability	×	√
large-scale deployment	×	√

Effectiveness. Table 1 shows the comparison of productivity indicators between human and DuIVRS. We observe that: (1) *TSR* of DuIVRS is 79.9%, which reaches 89.2% of human performance; (2) the cost per call is less than ¥0.2, which is only about 1/7 of the human cost; (3) although the current number of calls per day of DuIVRS was about 200,000, it has the ability to make unlimited calls daily. By contrast, a high-performance call center worker can only make up to 200 telephone calls a day; (4) the acquisition processes of DuIVRS are sufficiently standardized. By contrast, it is difficult to standardize manual calls; (5) the acquisition quality of DuIVRS is highly stable. By contrast, the highly repetitive and monotonous nature of call center work inevitably makes it difficult to consistently yield high-quality results; and (6) the advantages of low cost, high yield, standardization, and stability enable practical large-scale deployment of DuIVRS. The results demonstrate that DuIVRS is an industrial-grade and robust solution for cost-effective, large-scale acquisition of POI attributes.

Scalability. DuIVRS exhibits promising scalability in practice, making it advantageous to adapt to new attribute acquisition tasks. For example, during the fourth quarter of 2021, we extended DuIVRS to acquire parking attributes of businesses. This task is first decomposed into five subtasks, including: ask name, ask if parking lot is available, ask if parking lot is public, ask if parking lot is free, and ask parking price. Then, the dialogue structure and process are rapidly designed and completed through the dialogue customization and management platform. The process module is deployed into DuIVRS within one day, without any code to implement it. After launching, the feedback loop strategy is applied to iteratively improve its dialogue success rates. For now, DuIVRS is able to solidly perform acquisition tasks for 14 types of POI attributes.

Failure Analysis. Table 2 shows the distribution of reasons in 200 failure cases. First, 38.0% of cases are induced by NLU, which suggests that it is necessary to substantially improve NLU. For example, we can adopt the geography-and-language pre-trained model ERNIE-GeoL [15] to improve NLU. Second, 8.5% of cases are induced by ASR errors, and 8.5% of cases are caused by a machine-sounding voice. This suggests that the performance of ASR and TTS plays an important role in a telephonic IVR system. Third, 47.5% of cases are caused by factors beyond the control of DuIVRS, such as people have no time or no idea how to provide information.

Table 2: The distribution of reasons in 200 failure cases.

Reason	# of Cases (%)
people are uncertain about POI attributes	**33 (16.5%)**
people actively terminate the call	**125 (62.5%)**
low tolerance to machine-sounding voice	17 (8.5%)
people have no time to provide information	26 (13.0%)
DuIVRS makes no response to people	64 (32.0%)
intent is out of the pre-defined set	21 (10.5%)
intent is not recognized by NLU	43 (21.5%)
others	18 (9.0%)
failed to extract slot values in the voice	**29 (14.5%)**
ASR errors result in missing of keywords	17 (8.5%)
no ASR errors but still failed	12 (6.0%)
others	**13 (6.5%)**
total	**200 (100%)**

5 DISCUSSION

The deployment of DuIVRS has achieved four major benefits. (1) The tasks of determining the validity of existing POI attributes and filling in the missing POI attributes in a large-scale POI database are typically labor-intensive and costly. With the ability of cost-effective, large-scale POI attribute acquisition provided by DuIVRS, the productivity and efficiency can be significantly improved. (2) By using DuIVRS to automatically update POI attributes at Baidu Maps, we can provide users with up-to-date POI information to help them make informed decisions about local businesses. (3) To ensure potential customers find and learn more accurate information about local businesses online, business owners periodically update their business details (e.g., business status and business hours) on websites or Maps. With the assistance of DuIVRS, the efforts required for manually updating business information can be significantly reduced. Practical observation and analysis from a 3-year longitudinal study also show that business owners are generally willing to cooperate with DuIVRS to provide accurate POI attribute information. (4) When a public emergency event occurs, such as the COVID-19 pandemic, the businesses could be forced to dramatically change. Providing timely business changes can better serve the public interest. The high concurrency capability of DuIVRS enables us to accurately update business listings on Baidu Maps in very little time, which can help users suffer as little inconvenience as possible when visiting POIs that affect people's daily necessities of life (e.g., pharmacies, grocery stores, and restaurants).

Although business owners are generally willing to cooperate with DuIVRS, we still keep the individual well-being of business owners in mind, and we offer them maximum autonomy during the call. First, DuIVRS is scripted to clearly reveal its identity and purpose when a phone call is answered. During the call, business owners have the initiative to decide whether to continue the call. Second, we only use DuIVRS to initiate calls to businesses that are in conformity with the following two conditions: (1) POIs with missing attributes, and (2) POIs with abnormal attributes that are reported by users. Third, in order to minimize the disturbance to business owners, we strictly restrict the frequency of calls to the same business.

6 CONCLUSIONS

This paper suggests a production-proven solution for cost-effective, large-scale POI attribute acquisition via a telephonic IVR system. Since its deployment, DuIVRS has significantly improved the effectiveness of POI attribute acquisition at Baidu Maps.

REFERENCES

[1] Pranav Bhagat, Sachin Kumar Prajapati, and Aaditeshwar Seth. 2020. Initial Lessons from Building an IVR-Based Automated Question-Answering System. In *Proceedings of the 2020 International Conference on Information and Communication Technologies and Development*. Article 27, 5 pages.

[2] Nathanael Chambers and Dan Jurafsky. 2011. Template-based information extraction without the templates. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*. 976–986.

[3] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A Survey on Dialogue Systems: Recent Advances and New Frontiers. *ACM SIGKDD Explorations Newsletter* 19, 2 (2017), 25–35.

[4] Yudong Chen, Xin Wang, Miao Fan, Jizhou Huang, Shengwen Yang, and Wenwu Zhu. 2021. Curriculum meta-learning for next POI recommendation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2692–2702.

[5] Hsiu-Min Chuang and Chia-Hui Chang. 2015. Verification of POI and Location Pairs via Weakly Labeled Web Data. In *Proceedings of the 24th International Conference on World Wide Web*. 743–748.

[6] Don A Dillman, Glenn Phelps, Robert Tortora, Karen Swift, Julie Kohrell, Jodi Berck, and Benjamin L Messer. 2009. Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Social science research* 38, 1 (2009), 1–18.

[7] Miao Fan, Jizhou Huang, and Haifeng Wang. 2022. DuMapper: Towards Automatic Verification of Large-Scale POIs with Street Views at Baidu Maps. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*.

[8] Miao Fan, Yibo Sun, Jizhou Huang, Haifeng Wang, and Ying Li. 2021. Meta-Learned Spatial-Temporal POI Auto-Completion for the Search Engine at Baidu Maps. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2822–2830.

[9] Li Gong and Jennifer Lai. 2003. To Mix or Not to Mix Synthetic Speech and Human Speech? Contrasting Impact on Judge-Rated Task Performance versus Self-Rated Performance and Attitudinal Responses. *International Journal of Speech Technology* 6, 2 (2003), 123–131.

[10] Donghoon Ham, J-G Lee, Youngsoo Jang, and K-E Kim. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 583–592.

[11] Jizhou Huang, Haifeng Wang, Shiqiang Ding, and Shaolei Wang. 2022. DuIVA: An Intelligent Voice Assistant for Hands-free and Eyes-free Voice Interaction with the Baidu Maps App. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3040–3050.

[12] Jizhou Huang, Haifeng Wang, Miao Fan, An Zhuo, and Ying Li. 2020. Personalized Prefix Embedding for POI Auto-Completion in the Search Engine of Baidu Maps. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2677–2685.

[13] Jizhou Huang, Haifeng Wang, Miao Fan, An Zhuo, Yibo Sun, and Ying Li. 2020. Understanding the Impact of the COVID-19 Pandemic on Transportation-Related Behaviors with Human Mobility Data. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3443–3450.

[14] Jizhou Huang, Haifeng Wang, Yibo Sun, Miao Fan, Zhengjie Huang, Chunyuan Yuan, and Yawen Li. 2021. HGAMN: Heterogeneous Graph Attention Matching Network for Multilingual POI Retrieval at Baidu Maps. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 3032–3040.

[15] Jizhou Huang, Haifeng Wang, Yibo Sun, Yunsheng Shi, Zhengjie Huang, An Zhuo, and Shikun Feng. 2022. ERNIE-GeoL: A Geography-and-Language Pre-Trained Model and Its Applications in Baidu Maps. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3029–3039.

[16] Jizhou Huang, Ming Zhou, and Dan Yang. 2007. Extracting Chatbot Knowledge from Online Discussion Forums. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence*. 423–428.

[17] N Kaji, M Okamoto, and S Kurohashi. 2004. Paraphrasing Predicates from Written Language to Spoken Language Using the Web. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*. 241–248.

[18] Chin-Hui Lee, Bob Carpenter, Wu Chou, Jennifer Chu-Carroll, Wolfgang Reichl, Antoine Saad, and Qiru Zhou. 2000. On natural language call routing. *Speech Communication* 31, 4 (2000), 309–320.

[19] Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1437–1447.

[20] Yaniv Leviathan and Yossi Matias. 2018. Google Duplex: An AI system for accomplishing real-world tasks over the phone. Google AI blog.

[21] Olivier Pietquin and Thierry Dutoit. 2006. A Probabilistic Framework for Dialog Simulation and Optimal Strategy Learning. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 2 (2006), 589–599.

[22] Adam Rae, Vanessa Murdock, Adrian Popescu, and Hugues Bouchard. 2012. Mining the Web for Points of Interest. In *The 35th International ACM SIGIR conference on research and development in Information Retrieval*. 711–720.

[23] Jérôme Revaud, Matthijs Douze, and Cordelia Schmid. 2012. Correlation-Based Burstiness for Logo Retrieval. In *Proceedings of the 20th ACM Multimedia Conference*. 965–968.

[24] Hang Su, Shaogang Gong, and Xiatian Zhu. 2017. WebLogo-2M: Scalable Logo Detection by Deep Learning from the Web. In *2017 IEEE International Conference on Computer Vision Workshops*. 270–279.

[25] B Suhm and P Peterson. 2002. A data-driven methodology for evaluating and optimizing call center IVRs. *Internat. J. of Speech Technology* 5, 1 (2002), 23–37.

[26] Yibo Sun, Jizhou Huang, Chunyuan Yuan, Miao Fan, Haifeng Wang, Ming Liu, and Bing Qin. 2021. GEDIT: Geographic-Enhanced and Dependency-Guided Tagging for Joint POI and Accessibility Extraction at Baidu Maps. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 4135–4144.

[27] Chris van der Lee, Emiel Krahmer, and Sander Wubben. 2018. Automated learning of templates for data-to-text generation: comparing rule-based, statistical and neural methods. In *Proceedings of the 11th International Conference on Natural Language Generation*. 35–45.

[28] Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A Network-based End-to-End Trainable Task-oriented Dialogue System. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 438–449.

[29] Michael Witbrock, David Baxter, Jon Curtis, Dave Schneider, Robert Kahlert, Pierluigi Miraglia, Peter Wagner, Kathy Panton, Gavin Matthews, and Amanda Vizedom. 2003. An Interactive Dialogue System for Knowledge Acquisition in Cyc. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*. 138–145.

[30] Congxi Xiao, Jingbo Zhou, Jizhou Huang, An Zhuo, Ji Liu, Haoyi Xiong, and Dejing Dou. 2021. C-watcher: A framework for early detection of high-risk neighborhoods ahead of COVID-19 outbreak. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 4892–4900.

[31] Canwen Xu, Jing Li, Xiangyang Luo, Jiaxin Pei, Chenliang Li, and Donghong Ji. 2019. DLocRL: A Deep Learning Pipeline for Fine-Grained Location Recognition and Linking in Tweets. In *The World Wide Web Conference*. 3391–3397.

[32] Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. 2017. Building Task-Oriented Dialogue Systems for Online Shopping. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. 4618–4625.

[33] Xuejie Zhang, Samarth Agarwal, Ruth Choy, Kay Jan Wong, Lecia Lim, Ying Yang Lee, and John Jianan Lu. 2020. Personalized Digital Customer Services for Consumer Banking Call Centre using Neural Networks. In *2020 International Joint Conference on Neural Networks (IJCNN)*. 1–7.