IRWoZ: Constructing An Industrial Robot Wizard-of-OZ Dialoguing Dataset

LI, Chen; Chrysostomou, Dimitrios; Zhang, Xiaochun; Yang, Hongji

Published in:
IEEE Access

DOI (link to publication from Publisher):
10.1109/ACCESS.2023.3259325

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 20 January 2023, accepted 9 March 2023, date of publication 20 March 2023, date of current version 24 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3259325

RESEARCH ARTICLE

IRWoZ: Constructing an Industrial Robot Wizard-of-OZ Dialoguing Dataset

CHEN LI[1], (Member, IEEE), DIMITRIOS CHRYSOSTOMOU[1], (Member, IEEE), XIAOCHUN ZHANG[2], (Member, IEEE), AND HONGJI YANG[3]

[1]Department of Materials and Production, Aalborg University, 9220 Aalborg, Denmark
[2]School of Management Science and Computer, Anhui University of Finance and Economics, Bengbu, Anhui 233030, China
[3]School of Computing and Mathematical Sciences, University of Leicester, LE1 7RH Leicester, U.K.

Corresponding author: Chen Li (cl@mp.aau.dk)

This work was supported in part by the Project of the Natural Science Foundation of the Education Department of Anhui Province under Grant KJ2020A0012, in part by the Program of School Scientific Research of the Anhui University of Finance and Economics under Grant ACKYC22092, in part by the European Union (EU's) SMART EUREKA Program under Grant S0218-chARmER, in part by the Innovation Fund Denmark under Grant 9118-00001B, and in part by the H2020-WIDESPREAD Project (Networking for Research and Development of Human Interactive and Sensitive Robotics Taking Advantage of Additive Manufacturing–R2P2) under Grant 857061.

ABSTRACT Enabling a flexible and natural human-robot interaction (HRI) for industrial robots is a critical yet challenging task that can be facilitated by the use of conversational artificial intelligence (AI). Prior research has concentrated on strengthening interactions through the deployment of social robots, while disregarding the capabilities required to boost the flexibility and user experience associated with human-robot collaboration (HRC) on manufacturing tasks. One of the main challenges is the lack of publicly available industrial-oriented dialogue datasets for the training of conversational AI. In this work, we present an Industrial Robot Wizard-of-Oz Dialoguing Dataset (IRWoZ) focused on enabling HRC in manufacturing tasks. The dataset covers four domains: assembly, transportation, position, and relocation. It is created using the Wizard-of-Oz technique to be less noisy. We manually constructed, annotated and validated dialogue segments (e.g., intentions, slots, annotations), as well as the responses. Building upon the proposed dataset, we benchmark it on the state-of-the-art (SoTA) language models, generative pre-trained (GPT-2) models, on dialogue state tracking and response generation tasks. We expect that the IRWoZ dataset will facilitate exciting ongoing dialogue research and we provide it freely accessible at https://github.com/lcroy/ToD4IR/tree/main/dataset.

INDEX TERMS Data collection, data annotation, dialogue systems, virtual assistants, human-robot interaction.

I. INTRODUCTION

Since the 1969, the first industrial robot, Unimate,[1] was manufactured and deployed, industrial robots have been designed and utilized in production extensively for operations such as welding [1], bin picking [2] and assembly [3], [4], [5]. Various technologies, e.g., computer vision and force sensing, have been introduced to facilitate human-robot interaction (HRI) for manufacturing tasks [6]. Furthermore, with the development of artificial intelligence (AI), the AI-empowered HRI for industrial robots is booming in industrial environments worldwide. To increase the overall performance, manufacturers are investigating intelligent and user-friendly interfaces to prevent steep learning curves and enhance the operator's experience when collaborating with robots [7]. Robots are expected to be capable of interpreting operators' intent and naturally interacting with the operators.

Various research work has been proposed in this direction, with the focus on speech-enabled virtual assistants (VAs) [8], [9], [10], [11]. Other commercial products, e.g., Alexa,[2]

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Liu.

[1]https://www.automate.org/a3-content/joseph-engelberger-unimate

[2]https://developer.amazon.com/en-US/alexa

Google Assistant,[3] Siri,[4] are widely used for entertainment or personal purpose. However, only few studies employ natural language interfaces in industrial and manufacturing environments while using industrial robots [7], [12], [13], [14], [15], [16], [17]. A highlighted challenge in these studies is the lack of a dialogue corpus tailored for such VAs for manufacturing purposes. Therefore, the creation of such dialogue dataset can benefit both dialogue research community and industrial companies for carrying on the the research and development of speech-enabled VA in the industrial setup for HRI with industrial robots.

This paper proposes an industrial robots Wizard-of-OZ dialogue (IRWoZ) dataset, a multi-turn conversational dialogue corpus spanning across four domains of industrial robots. Each dialogue includes the user's utterance, annotated belief states, system action and system response. Furthermore, the system response is composed of task-related response and small talk response to form more humanized replies. Therefore, IRWoZ can be used to develop task-oriented dialogue systems for industrial tasks and serve as a new benchmark for dialogue research community and industrial companies. The IRWoZ contains 401 dialogues, and follows a similar data structure as the most popular dialogue corpus MultiWoZ [18], which can be easily extended to other domains and reused for training various existing dialogue systems, e.g., [19] and [20].

The remainder of this work is structured as follows: Section II summarizes the related work. Section III presents the data collection method of IRWoZ. Section IV describes the analysis of the dialogue corpus. To show the potential and usefulness of the proposed IRWoZ, we benchmark Generative Pre-trained (GPT) models on the IRWoZ and report the evaluation results of perplexity, bilingual evaluation understudy (BLEU), joint goal accuracy, and slot accuracy in section V. Last, Section VII concludes the paper.

II. RELATED WORK

Due to the fact that the quality of dialogue datasets affects the performance of human-machine dialogue systems, dataset construction has constituted an active area of research for decades [21]. Existing dialogue systems can be roughly classified into two types: task-oriented and open domain dialogue systems [22].

Initial task-oriented dialogue datasets mainly focus on single task, i.e, one domain. As an example, the ATIS dataset is focused in the task of inquiring flight information, and it consists of 17 categories, including 4,978 and 893 utterances for training and test set respectively [23]. The Dialogue State Tracking Challenge, DSTC2 [24], is a single domain dataset aimed on restaurant booking for users. DSTC2 includes 1612, 506 and 1116 user utterances for training, validation and test set, respectively. WOZ2.0 consists of 1200 dialogues for the restaurant domain [25]. Due to user conversations often change domains from one to another, Multi-WoZs modified the WoZ framework to make it suitable for multi domains where a total of 10k dialogues were collected [18]. KVRET is also a multi-domains corpus [26], including calendar scheduling, weather information retrieval and PoI navigation field. It includes 2425, 302, 304 utterances for training, validation and test set, respectively. KdConv contains 4,5000 Chinese dialogues from film, music and travel domains, and an average dialogue turn of 19, with a total of 86,000 sentences [27]. These dialogues include in-depth discussions on relevant topics and natural transitions between multiple topics.

Those datasets are mainly used for training task-oriented dialogue (ToD) systems. To accomplish task and improve accuracy, Li et al. [28] suggested a neural dialogue system that works in end2end manner. To improve performance on downstream tasks, such as response generation, Wu et al. presented a ToD-BERT that is pre-trained to mimic conversation behavior [29]. Minimalist Transfer Learning (MinTL) was developed to enhance end-to-end response generation and was distinguished from TOD-BERT by the use of a copy mechanism to inject the prior dialogue states into the current one [30]. Using transfer learning from a pre-trained language model based on open-domain, Hosseini-Asl et al. [19] proposed SimpleTOD increased the performance of the dialogue model by treating the whole ToD as a single sequence prediction issue. To better extract information from both utterances and graphs, Chen et al. proposed a graph attention network and used a recurrent graph attention network to manage state updates [31]. Liu et al. also considered human input during the training phase of the end-to-end model, they leveraged human input to boost system performance [32]. To improve its task success rate, Soloist [20] employed task-grounded pre-training to learn tasks while benefiting from a cheap annotation cost for the training dataset. An end-to-end differentiable KB-Infobot was introduced by Dhingra et al., which enhanced the system's reliability and allowed for more varied inquiry formats [33]. The large pre-trained language model is leveraged by the suggested Alternating Roles Dialog Model (ARDM) [34].

Compared to task-oriented, open domain datasets contains daily life topics about emotion, mood or just common small talks. MCTest [35] is a dataset of 500 stories and 2000 questions. DailyDialog consists of daily dialogues with emotion information written by human users. The corpus contains 13,118 dialogues, which has less noisy, more humanlike response [36]. Conversations are extracted from movie for the Cornell Movie-Dialogue Corpus [37] where there are many metadata included such as genre, release year, and IMDB rating in conversations with a corpus containing 300,000 utterances. NPS [38] Internet Chatroom Conversations dataset consists of 10,567 utterances from online chat between October and November in 2006. Each utterance includes some speech and dialogue act information. The Internet Relay Chat (IRC) [39] Corpus comes from IRC chat logs. It contains approximately 50 hours of chat, with

[3]https://developers.google.com/assistant

[4]https://developer.apple.com/siri/

an estimated 20,000 utterances from the IRC. This dataset includes technical conversations and occasional social chats and there are approximately 1,500 utterances with annotated ground-truth conversations.

OpenViDial [40] contains a total number of 1.1 million text and visual contexts stored in images, considering visual and text contexts at the same time. OpenDomain Spoken Question Answering Dataset (ODSQA) dataset is an open domain spoken question answering dataset, which includes 3,000 questions [41]. ConvAI2 with 20,000 dialogues about daily life topics, is an extended version of the Persona-Chat dataset [42], including the training, validation and test sets consisting of 17,878, 1,000 and 1,015 dialogues respectively. Coached Conversational Preference Elicitation (CCPE) contains 502 dialogues discussing movie preferences. The dataset was collected based on Wizard-of-Oz method [43]. Wizard of Wikipedia is a large dataset containing retrieved knowledge from Wikipedia. The dataset consists of 22,311 dialogues, dividing into 166,787 for training, 17,715 for validation, and 17,497 dialogues for test [44].

Despite the fact that various domains have been investigated either through the task-oriented manner or open domain, none of the public available dialogue datasets are collected for building dialogue systems for HRI in industrial robots for manufacturing tasks. To the best of our knowledge, IRWoZ is the first dialogue dataset with a focus on collecting and annotating dialogue corpus on four popular manufacturing tasks e.g., delivery, position, assembly, and relocation for industrial robots.

III. DATA COLLECTION METHOD

In this section, we investigate the robot skills for manufacturing to identify the industrial robot's capability on the cluster level. The manufacturing tasks, which are suitable for speech-enabled HRI, are chosen to define the boundaries of the dialogue datasets. Second, we explain how to use the WoZ approach to simulate the dialogue between the shop floor worker and the robots. Conversation strategies are leveraged to boost the user experience by provide the hybrid responses, task-related response and the small talk response.

A. INDUSTRIAL TASKS

Robot-based production is becoming the mainstream of industrial manufacturing [45]. In general, mobile robots, e.g., mobile industrial robots (MiR),[5] and industrial manipulator, e.g., Universal Robots (UR),[6] are most common industrial robots for building the robot-based production. To introduce the dialogue system to those robots, the following questions need to be answered before the data collection.

- Q1: What are the general task categories of industrial robots and in which domains do they operate?
- Q2: What are the suitable task(s) for speech-enabled HRI with industrial robots?

Various approaches has been proposed to answer the Q1. In [46], a taxonomic framework is presented for task modeling and knowledge transfer in manufacturing robotics. It decomposes the tasks into skills, e.g., detect, fasten, and coordinate skill. Wordcloud is also used to represent the occurrence of words in the robot's tasks, skills and primitives, e.g., button-press, clean-wipe, placement-pick [47]. Segura et al. classify the manufacturing tasks into four categories, assembly, material manipulation, machining processes, and quality inspection [48]. Our work is inspired by [45] which identifies nine general task categories in three domains, i.e., logistic tasks (transportation, part feeding (multi), part feeding (single)), assistive tasks (machine tending, assembly, inspection, process) and service tasks (maintenance, repair and overhaul, cleaning). It provides a clear three-layer architecture of skills, tasks and primitives.

In our work, four tasks are selected to answer the second question. Within the Logistics domain, transportation often involves the internal *Delivery* task in which a mobile industrial robot needs to deliver goods to a desired position. To enable such functions, the mobile industrial robot is also required to be able to mark *Position* on the digital map or build a path to to reach a position. Example dialogues can be *"hey robot, can you deliver this box to the warehouse?"* or *"Robot, please move to the production cell."*. Another two tasks are identified within the Assistive domain: *Assembly* and *Relocation*. Operator may send a request to an industrial manipulator, "Start to assemble product" or "Bring the screwdriver.".

B. DIALOGUE SIMULATION: WoZ APPROACH

The WoZ approach (i.e., human-to-human method) is used in this study to mimic dialogue between shop floor worker and industrial robots. To assist the process of collecting the dialogue corpus, a web application built on the Flask[7] web framework is designed and implemented.[8] The overall architecture of the web application of the IRWoZ dialogue simulation framework is depicted in Figure 1. It follows the Client-Server style architecture. The client side comprises user and wizard interfaces, while the server side includes the dialogue controller, belief state verification, dialogue act generation, dialogue annotation, and dialogue autosaving. The ground truth data (such as employee IDs, position names, and product categories) of the four chosen domains are stored in a domain-specific database that has five tables: Employee, Area_Location, Position, Product, and Object. Such a database is utilized for dialogue act generation and belief state verification. The simulated conversations will be automatically saved in the IRWOZ database.

The dialogue simulation preparation is run in four steps to obtain an high quality and less noises dialogues.

- *Step one: Participant selection*. The invited participants are from both academia (with background of Robotics

[5] https://www.mobile-industrial-robots.com/
[6] https://www.universal-robots.com/da/
[7] https://flask.palletsprojects.com/en/2.0.x/
[8] https://github.com/lcroy/ToD4IR/tree/main/IRWoZ_interface

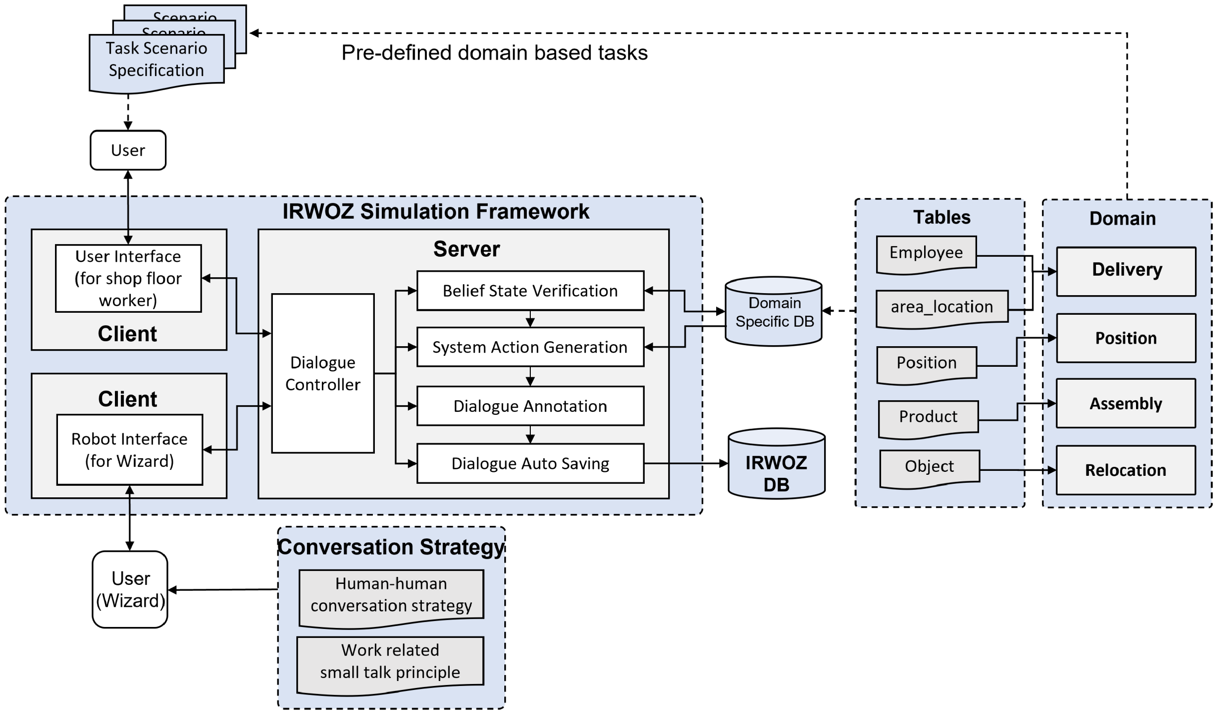


FIGURE 1. The overall architecture of the dialogue simulation framework for building the IRWoZ dataset.

and Automation, Computer Science, and Culture and Linguistic) and industry (shop floor worker and factory engineers). The majority of participants have knowledge, skills, and experience within the area of manipulation with industrial robots.

- *Step two: Introduction of Industrial robots and task.* The participants are introduced to industrial robots and the tasks associated with the specific domains. As previously stated, the MiR 200 and Franka Emika robots are the primary robots in our scenario. The participants acquired knowledge of the domain's most critical activities (e.g., instruct MIR to transport a package to a location) through the use of the aforementioned robots.
- *Step three: Watching video demonstrations.* The participants are invited to watch the pre-recorded video demos based on our prior work with the industrial intelligent virtual assistant [7], [12] in order to understand the dialogues pattern between the human and the robots. The videos demonstrate language-assisted HRI through the dialogues with various industrial equipment and robots, including Lego counting machine,[9] MIR[10] and Franka Emika robot.[11] The dialogues cover the topics of production control, internal transportation and material grasping.
- *Step four: Instruction of IRWoZ Web application.* The participants are instructed on how to conduct a dialogue using the IRWoZ online application, including picking a role in the dialogue and the domain in which the dialogue will take place. The following sections go into further details.

[9]shorturl.at/dgvOR
[10]shorturl.at/lqCI4
[11]shorturl.at/aizY9

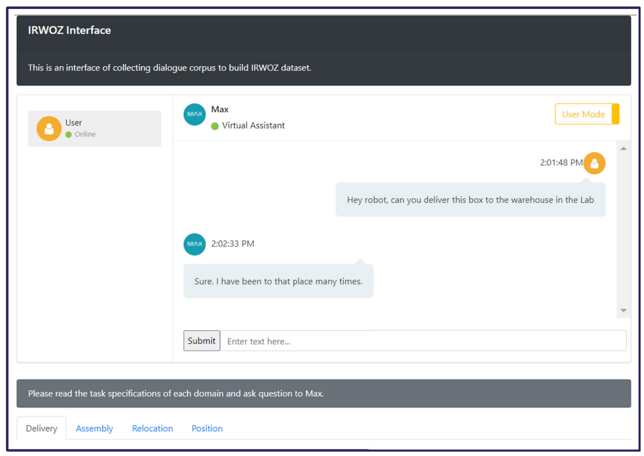


FIGURE 2. The graphical user interface. The top right button shows that the interface is in user mode at the moment. At the bottom, four buttons symbolize four domains. The user may type a message in the text field and then click the submit button to send it to the dialogue controller.

C. USER INTERFACE

Each dialogue involves the pairing of two participants. The participant interacting with the *User Interface* takes on the role of a shop floor worker (see Fig. 2). The participant is asked to randomly select a manufacturing task with a detailed

Scenario Name: move to a position

Precondition:
- The participant received the document of domain – Position
- The participant received the instruction of mobile robots
- The participant watched the online video of sample dialogue between the operator and robots
- The participant learned how to use IRWoZ application.

Steps:

1. The participant confirms that the interface is under "user mode"
2. The participant chooses the domain – "position" from the interface
3. The participant check the list of the positions from the document and perform either one of the following actions:
   - Enter a message in the input box and command robot to move to a previously defined position (Participant chooses the one from the list)
   - Enter a message in the input box and command robot to move to an undefined position (Participant names a position which is not on the list)
4. The participant may take the following actions in response to the system's reply.
   - Proceed to step 3
   - End the dialogue

FIGURE 3. A sample scenario specification of a position task.

task scenario specification (see Fig. 3). At the same time, the participant receives the following information of each domain:

- *Delivery*: A document containing a list of workers' names and locations of the factory. To accomplish a transportation job, users must inform the robot (wizard) not only of the delivery's destination, but also of the recipient's name. As a result, the participant can pick the recipient from a list of workers' names and the place to which the package should be delivered. Apart from the fundamental information necessary for a delivery task, we aim to cover a broad range of delivery task variations. As a result, users may choose not only the name of the goods they want to deliver but also its size and color during the conversation. The participant may additionally provide the sender's name if desired, such as *"Could you please deliver this little yellow screw driver to the warehouse, robot? and give it to John and inform him that it is from Martin."*
- *Position*: A digital map of factory with the positions marked. There are two types of the tasks, adding a position on the digital map and moving to a position, are currently supported by MIR 200. The participant can instruct the robot to move to the specific position if it is registered in the system, such as *"Hey robot, can you run down to the product cell one?"*. Additionally, adding a position to the digital map is supported only if the system is given with a unique position name.
- *Assembly*: a document containing a list of the products. The participant is informed what kind of the products can be assembled by Franka Emika robot. Participant can command the robot to assemble a product during the dialog. To add variation to the task, the user may additionally provide the quantity of desired product and the deadline of the assembly task, such as *"Hi robot, I need 10 white smart phones before 3pm."*.
- *Relocation*: a document containing a list of the materials. The participant may command the Franka Emika robot to grasp the required material on the work space if the object is listed on the document. To add variation to the task, the user may additionally provide the relative or specific location, such as *"Hello robot, can you take the PCB from the left side and put it in the yellow box?"*.

The maximum conversational depth and the number of continuous dialogue turns for a task are not limited to creating a natural dialogue environment [49]. Given that individuals, even when assigned the same task, structure their utterances differently, the shop floor worker is encouraged to compose the dialogue in their own unique manner. Though English is required as common working language for building the IRWoZ, some intentional grammar mistakes and various sentence structures are observed during the simulation. However, we intent to keep those as "white noise" in the dialogues since they create more natural real-world scenarios.

D. WIZARD INTERFACE

In comparison to the user, the Wizard, as an industrial robot, must be able to recognize the user's intent (e.g., the task the user requested) and confirm that the user provided all essential information to complete the task.

We create the *Wizard Interface* (see Fig. 4) that provides in-dialog instructions to assist wizards. Each domain has two distinct panels for describing the mandatory and optional information for completing a task. The wizard must first select the corresponding domain according to the message received from the user. Following that, the wizard needs to extract the core information, including the database and task-related information, and fill them in the corresponding panel. To assist the wizard in this process, each panel specifies the type of information that is required, for example, *"Ask user which location the box should be delivered."* for a delivery task. Such pre-defined in-dialog instructions guide wizards on how to collect and validate data from user's messages. Wizards do not only get inputs the from the *User Interface*, but also from the dialogue controller, which assists in verifying belief states, generating system actions, annotating the dialogue, and auto-saving the dialogue. Wizards are requested to provide two types of response: task-related response and small talk response (see section III-D1).

The graphical user interface and predefined conversation instructions work together to help wizards build consistent dialogues. This setting allows wizards to respond to unexpected user utterances.

1) CONVERSATION STRATEGY

In a task-oriented conversation, the human often steers the conversation by requesting information, whereas the robot

TABLE 1. The "ARE" principles [50] and three conversation strategies [49] for generating humanised response with examples.

Principle	Definition	Examples	Strategy	Definition	Examples
Anchor	the conversation with a topic that is part of speaker's mutual shared reality	*" Yes, I knew that place. It is a quite busy area."*	End topics with a suggestion	provide task-related suggestions	*Well, I don't know this place. Can you register it in the system first?*
Reveal	say something that has more information about you (using the Anchors mentioned above).	*" Well, it is quite new to me. I have not learned how to do that."*	Elicit more information	Invite operator to provide more information.	*Can you describe a bit more of the thing you want to assemble?*
Encourage	invite others to speak with a question	*" Hey, how are you? Do you have a nice day?"*	Clarifying	attempt to understand what the speakers express in their messages	*Do you mean you want to abort the task?*

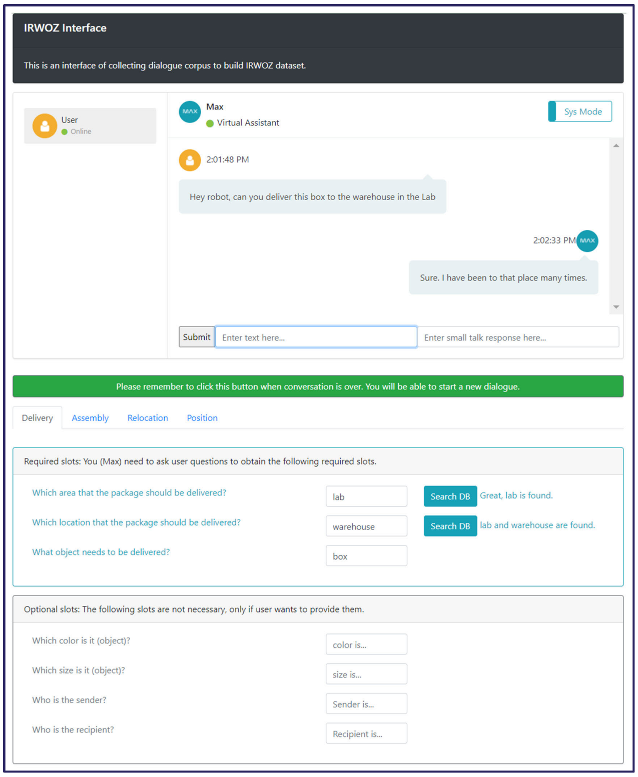


FIGURE 4. The wizard's graphical user interface. The top right button indicates that the interface is now in system mode. According to the user's request, the wizard can switch between up to four domains. There are two distinct panels for required and optional slots. If the slot requires database verification, the *Search DB* is presented. When the wizard recognizes the conclusion of the dialogue, the green button, in the center, needs to be clicked to automatically save it to the IRWoZ database. Two input text boxes are provided for the wizard to give task-related response and small talk response.

follows the user's lead and responds appropriately. However, such interaction typically results in increased efficiency at the cost of user engagement. For example, in a question "hello, robot, can you move to the warehouse now?", the robot might propose "Sorry, you may need to register the warehouse position in the system first.", instead of responding "No, I am unable to do that" to a user's request. Additionally, the response "Yes, I am pleased to assist." is more pleasant for the user than the response "Yes." in a query such as "Hi robot, assemble ten phones."

Therefore, three human-human conversation strategies [49] are introduced to boost user engagement: *end topics with a suggestion*, *elicit more information* and *clarifying*, in order to increase task completion rate while keeping a more humanised and task-related answer. Furthermore, inspired by [50], Anchor, Reveal and Encourage (ARE) principles are introduced to assist the wizard for small talk responses if it is desired. Table 1 explains the conversation strategies and small talk principles with examples [51].

IV. DIALOGUE ANNOTATION AND PROCEDURE

Participants are divided into two types of groups: Annotators and Validators. They are asked to complete the following two tasks, respectively.

- Task 1: Annotating the domain, slots, task-related responses, and small talk response of the dialogue corpus.
- Task 2: Labelling the responses as appropriate and inappropriate.

For task 1, the web application (see Figure 4) is leveraged to guide the annotation process. It provides the instructions on what are the required and optional database-related and task-related slots, and which domain is requested by the user. *Annotators* are asked to extract such information from the user's utterance and fill it into the application. Additionally, the application will automatically wrap it into a query string to verify the filled slots against the database if the slots are database related.

Inspired by the [52], we ask *Validators* to choose the appropriate and inappropriate dialogue corpus. Though leveraging the human-to-human method can generate more natural and humanised dialogues, inappropriate responses are observed. Validators are asked to use the following justifications for choosing an inappropriate candidate:

- Mislabelling: The responses are not correctly annotated (e.g., a task-related response is annotated as a small talk response, or the other way around). This is the most commonly discovered inappropriate case during the data annotation. For example, *"Sorry, I don't know how to assemble that product."* was mis-annotated as a task response: *"Sorry"*, and small talk response: *"I don't know how to assemble that product."* A task-related

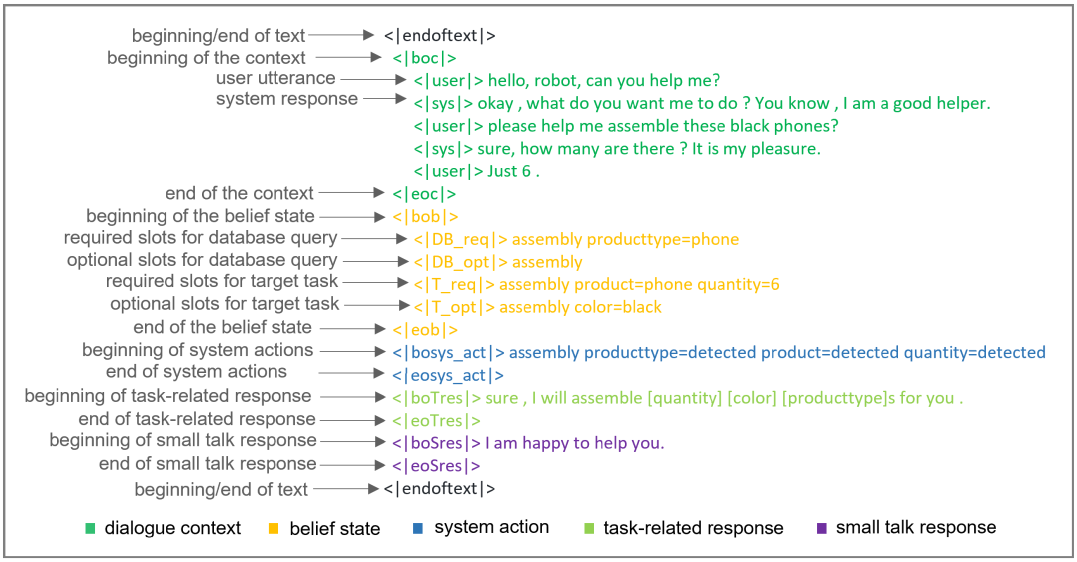


FIGURE 5. A dialogue example marked with special tokens.

response, however, should include sufficient information so that the end user is not left with an incomplete understanding. The simple response: *"Sorry"* does not explain why the robot is unable to execute the requested task in the absence of the statement *"I don't know how to assemble that product."* On the other hand, the purpose of small talk response is to make the conversation smooth and natural, not carry any task-related information. A response, *"Sure, I can do that. I have been to that area many times."*, for a requested *delivery* task should be annotated as task response *"Sure, I can do that."* and small talk response *"I have been to that area many times."*

- Overly humanized response: Despite the fact that the humanized response is one of the primary features of proposed datasets, the robot's response should be not overly emotional. For example, the response, "Ok, I will go to that position, but I don't like it there.", is identified as overly humanized response because the phrase *"but I don't like there"* attempts to express emotion and feeling, which could lead to confusion for a human worker.

In addition to the above two criteria, ARE principles and the conversation strategies (see Table 1) are printed and distributed to the annotators and validators in order to acquire high-quality responses. Inspired by [18], we conducted a two-phase trial to improve the performance of annotators and validators. In the first round, we randomly selected 10% of the dialogue corpus as test samples, which we distributed to the annotators. The annotated dialogue corpus was verified by validators afterwards to filter out the inappropriate responses. Annotators and validators were invited to participate in a follow-up meeting to discuss all filtered corpora. They were allowed to begin annotating and validating the whole dataset after passing the first trial.

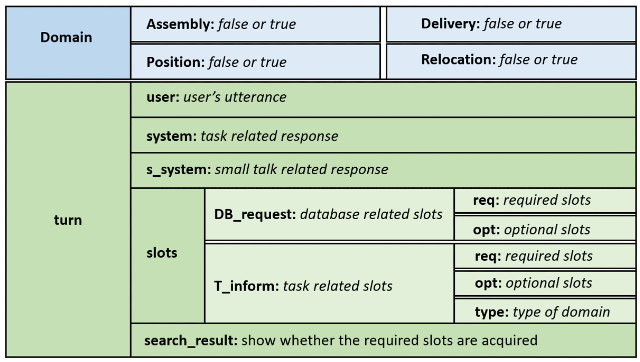


FIGURE 6. Data structure of the dialogue of IRWoZ dataset.

V. DIALOGUE ANALYSIS

Our IRWoZ has 401 dialogues that span across four domains. It is the first annotated task-oriented dialogue dataset made publicly available with a focus on industrial robots. The dataset has a similar structure as one of the most popular multi-domain datasets, MultiWOZ [53], [54], [55]. Additionally, another version of the IRWoZ dataset is also provided. The dialogues are marked with special tokens (see Figure 5), which may be beneficial to the dialogue research community and industrial companies for preparing the language model for dialogue system development.

In this section, we delve deeply into IRWoZ datasets, and show the basic features in the following aspects: data structure, data quality, and statistics.

A. DATA STRUCTURE

Due to the similar data structure to MultiWoZ, IRWoZ can be easily expanded, customized, benchmarked and evaluated for various dialogue systems. Table 2 presents the global

TABLE 2. The full ontology for all domains, including assembly, delivery, position and relocation, of IRWoZ dataset.

act type	Delivery_DB_request_req, Delivery_DB_request_opt Delivery_T_inform_req, Delivery_T_inform_opt Position_DB_request_req, Position_DB_request_opt Position_T_inform_req, Position_T_inform_opt Assembly_DB_request_req, Assembly_DB_request_opt Assembly_T_inform_req, Assembly_T_inform_opt Relocation_DB_request_req, Relocation_DB_request_opt Relocation_T_inform_req, Relocation_T_inform_opt Greet
dialogue slots	Delivery-area, Delivery-location Delivery-sender, Delivery-recipient Delivery-object, Delivery-color, Delivery-size Position-name, Position-operation Assembly-producttype, Assembly-product Assembly-quantity, Assembly-color, Assembly-style Relocation-object, Relocation-color, Relocation-size Relocation-from, Relocation-to

ontology with the list of considered acts type and dialogue slots.

Each dialogue of IRWoZ is formulated as a JavaScript Object Notation (JSON) string which includes two elements, *domain* and *turn* (see figure 6). The corresponding domain will be set up as "true" if the related dialogue is identified, for example, *"domain":{"assembly":true}*. Each *turn* is composed of multiple turns of dialogue between the user and the robot. The user's utterance, task-related response, and small talk response are annotated as the *"user"*, *"system"* and *"s_system"*, respectively. *DB_request* and *T_inform* are defined in *Slots* to list the database related and task related slots, separately. Both *DB_request* and *T_inform* contain the required and optional slots, *req* and *opt*. Furthermore, *type* is provided for defining the type of domain under the *T_inform*. *search_result* is included in the *turn* to indicate whether the required slots are obtained.

Appendices A and B provide examples of raw and marked dialogue corpus of each domain, respectively. The complete dataset can be accessed at https://github.com/lcroy/ToD4IR/tree/main/dataset.

B. DATA QUALITY

The dialogue corpus was acquired via a human-to-human approach, with the majority of participants being shop floor engineers, lab engineers, and researchers with expertise in production or robotics. In addition, a training session was organized to provide details of the simulation, including introduction to the robots and selected manufacturing tasks, demonstrating IRWoZ's web application, showing the dialogue examples, introducing the conversation strategy and small talk principles for response generation. This method

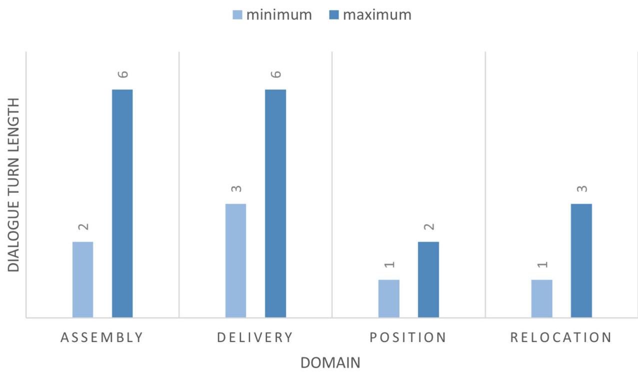


FIGURE 7. Dialogue turn distribution of each domain including the minimum and maximum dialogue turn length.

enabled us to acquire dialogue corpora that were cleaner and less noisy.

As mentioned above, data annotation leveraged a two-step process: annotating and validating. With the assistance of the IRWoZ web application, annotators are able to label the dialogue corpus with high accuracy. Additionally, validators helps to verify whether the annotated corpus is inappropriate. Finally, 22% of the corpus including user utterance and response was removed from the raw dialogues, i.e., 17 dialogues related to assembly, 46 related to delivery, 23 related to relocation and 27 dialogues related to the position task.

C. DATA STATISTICS

A total of 401 dialogues were collected and annotated after the identification and elimination of 113 (22%) inappropriate dialogues *turns*. Appendix C (figure 12, 13, 14 and 15) depicts the distribution of sentence length of user utterance, task-related response and small talk response of four tasks. The average length of the user utterance of each domain is around 35 tokens, ranging from 17 to 28 tokens for task-related response and 18 to 25 tokens for small talk response. As expected, each domain contains system responses without small talk involved. Similar to the [18], our responses are more varied, which improves the generalization ability for training model. Figure 7 shows the dialogue turn length distribution, including minimum and maximum turns, grouped by each domain. Due to the natural dialogue complexity of the *assembly* and *delivery* tasks, the maximum length of the turns for those two tasks is longer than the *position* and *relocation*. The minimum length of turns is one for both *position* and *relocation*.

VI. EXPERIMENTAL RESULTS

The proposed IRWoZ dataset provides a new benchmark for building dialogue systems, which focuses on HRI in four manufacturing tasks. To evaluate the IRWoZ, GPT based language models are leveraged as baseline models for testing the downstream dialogue tasks, dialogue state tracking, dialogue actions and response generation. The outcomes can serve as a

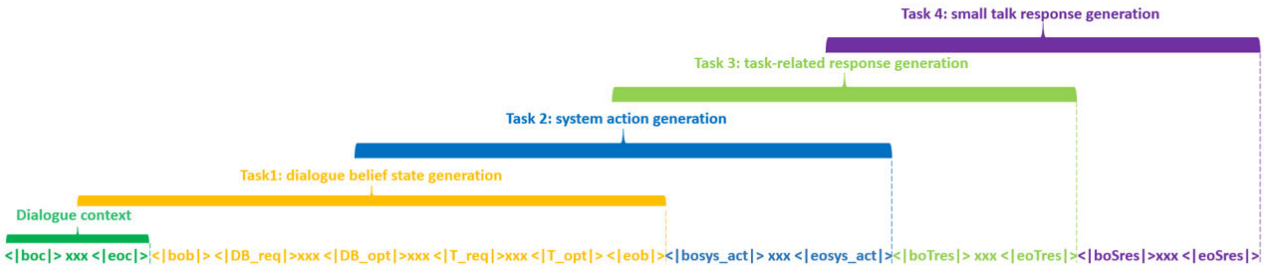


FIGURE 8. Dialogue tasks for evaluation of the IRWoZ based on GPT-2 models.

point of reference for development of dialogue systems which focus on similar manufacturing tasks.

A. TRAINING ON PRE-TRAINED LANGUAGE MODEL

We evaluated the IRWoZ dataset through an end-to-end approach by using three versions of SOTA GPT2 [56] language models, gpt2 (12-layer, 768-hidden, 12-heads, 117M parameters), gpt2-medium (24-layer, 1024-hidden, 16-heads, 345M parameters) and gpt2-large (36-layer, 1280-hidden, 20-heads, 774M parameters). They are trained on 40GB text of the web pages from outbound links on Reddit. All models work in an auto-regressive manner, i.e., they predict the next token after reading all of the preceding ones.

We divided the training process into four tasks (see figure 8). The initial input is the dialogue context, including the user's utterance and robot response. The models are trained to generate the belief state (see task 1 of figure 8) as the first step. Different from the belief state, the system action (see task 2 of figure 8) is not generated by the model but the query results of database (if the database-related slots are identified from the users' utterance) and task (if the task-related slots are identified). The model takes the dialogue context, belief state, and system action as input, and outputs the task-related and small talk responses sequentially as task 3 and task 4 (see task 3 and task 4 of figure 8).

B. EVALUATION

Four automatic metrics, joint goal accuracy (JGA) [57], slot accuracy (SA) [58], BLEU [59] and perplexity [60], are chosen for the evaluation.

- Joint goal accuracy. The output of the dialogue state tracker is compared to the ground truth label at the end of each discourse. The proportion of dialogue turns in which the value of each slot is correctly predicted is known as the joint goal accuracy.
- Slot accuracy. It compares each *(domain, slot, value)* triplet with the corresponding ground-truth label. Compared with the joint goal accuracy, its evaluation granularity is more refined.

TABLE 3. Performance comparison of three different model architectures on 30% of IRWoZ dataset.

Models \ Metrics	Perplexity	BLEU-1	BLEU-2	BLEU-3	BLEU-4	SA	JGA
gpt2	**1.08**	0.3636	0.2265	0.1522	0.1035	0.837	0.541
gpt2-medium	1.11	0.3807	0.2263	**0.2776**	0.1868	**0.942**	**0.795**
gpt2-large	1.11	**0.4167**	**0.3199**	0.273	**0.2356**	0.923	0.760

TABLE 4. Performance comparison of three different model architectures on 50% of IRWoZ dataset.

Models \ Metrics	Perplexity	BLEU-1	BLEU-2	BLEU-3	BLEU-4	SA	JGA
gpt2	**1.07**	0.3597	0.2394	0.1684	0.1178	0.896	0.658
gpt2-medium	1.08	**0.4928**	0.4048	0.3578	0.3191	**0.957**	**0.849**
gpt2-large	1.08	0.49	**0.4072**	**0.3639**	**0.3272**	0.95	0.815

TABLE 5. Performance comparison of three different model architectures on 70% of IRWoZ dataset.

Models \ Metrics	Perplexity	BLEU-1	BLEU-2	BLEU-3	BLEU-4	SA	JGA
gpt2	**1.06**	0.3563	0.2414	0.1764	0.1295	0.94	0.781
gpt2-medium	1.07	0.5256	0.4451	0.406	0.3738	0.962	0.863
gpt2-large	**1.06**	**0.5909**	**0.5186**	**0.4827**	**0.4523**	**0.969**	**0.884**

TABLE 6. Performance comparison of three different model architectures on 100% of IRWoZ dataset.

Models \ Metrics	Perplexity	BLEU-1	BLEU-2	BLEU-3	BLEU-4	SA	JGA
gpt2	**1.05**	0.3816	0.268	0.2088	0.1651	0.949	0.801
gpt2-medium	**1.05**	0.6214	0.562	0.5352	0.5114	0.979	0.925
gpt2-large	**1.05**	**0.6493**	**0.5949**	**0.5681**	**0.5446**	**0.983**	**0.938**

- BLEU. It evaluates how natural and fluency the replies are in comparison to those provided by human. There are also many variants of BLEU. According to n-gram (i.e., the number of consecutive words being n.),[12] it can be divided into various evaluation indicators. Common indicators are BLEU-1, BLEU-2, BLEU-3, and BLEU-4.
- Perplexity. It is algebraically equal to the inverse of the geometric mean per-word likelihood and monotonically decreases in the likelihood of the test data. As a consequence of this, a model that has shown better generalization performance would often have a lower perplexity score.

[12]https://en.wikipedia.org/wiki/N-gram

Domain	Raw Dialogue Samples
Delivery	"DL00101.json": {"domain": {"assembly": false,"delivery": true},"turn": [{"user": "Could you deliver the box for me please?","system": "sure, where do you want to deliver it?","s_system": "","slots": {"assembly": {"DB_request": {"req": {"producttype": ""},"opt": {}},"T_inform": {"req": {"product": "","quantity": ""},"opt": {"color": "","holeonPCB": "","deadline": ""},"type": ""}},"delivery": {"DB_request": {"req": {"area": "not_mentioned","location": "not_mentioned"},"opt": {"sender": "not_mentioned","recipient": "not_mentioned"}},"T_inform": {"req": {"object": "pens"},"opt": {"color": "not_mentioned","size": "not_mentioned"},"type": "delivery"}}},"search_result": {"area": "null","location": "null","object": "null"}},{"user": "Yes, it is office in the lab.","system": "OK, i found location office in the lab.","s_system": "I'm so happy! ","slots": {"assembly": {"DB_request": {"req": {"producttype": ""},"opt": {}},"T_inform": {"req": {"product": "","quantity": ""},"opt": {"color": "","holeonPCB": "","deadline": ""},"type": ""}},"delivery": {"DB_request": {"req": {"area": "lab","location": "office"},"opt": {"sender": "not_mentioned","recipient": "not_mentioned"}},"T_inform": {"req": {"object": "box"},"opt": {"color": "not_mentioned","size": "not_mentioned"},"type": "delivery"}}},"search_result": {"area": "detected","location": "detected","object": "null"}},{"user": "Great, please delivery it now","system": "Sure, I can execute the mission, but I need to know the recipient.","s_system": "Sorry for time","slots": {"assembly": {"DB_request": {"req": {"producttype": ""},"opt": {}},"T_inform": {"req": {"product": "","quantity": ""},"opt": {"color": "","holeonPCB": "","deadline": ""},"type": ""}},"delivery": {"DB_request": {"req": {"area": "lab","location": "office"},"opt": {"sender": "not_mentioned","recipient": "not_mentioned"}},"T_inform": {"req": {"object": "box"},"opt": {"color": "not_mentioned","size": "not_mentioned"},"type": "delivery"}}},"search_result": {"area": "detected","location": "detected","object": "null"}},{"user": "yes, send it to Tom please.","system": "Sure, can I do the mission now?","s_system": "If you would","slots": {"assembly": {"DB_request": {"req": {"producttype": ""},"opt": {}},"T_inform": {"req": {"product": "","quantity": ""},"opt": {"color": "","holeonPCB": "","deadline": ""},"type": ""}},"delivery": {"DB_request": {"req": {"area": "lab","location": "office"},"opt": {"sender": "not_mentioned","recipient": "Tom"}},"T_inform": {"req": {"object": "box"},"opt": {"color": "not_mentioned","size": "not_mentioned"},"type": "delivery"}}},"search_result": {"area": "detected","location": "detected","recipient": "detected","object": "null"}}]}
Assembly	"DL00119.json": {"domain": {"assembly": true,"delivery": false},"turn": [{"user": "How are you, friend! Can you assembly five black phones for me?","system": "Of course. When do you want me to do?","s_system": "I'm free now.","slots": {"assembly": {"DB_request": {"req": {"producttype": "not_mentioned"},"opt": {}},"T_inform": {"req": {"product": "phones","quantity": "five"},"opt": {"color": "black","holeonPCB": "not_mentioned","deadline": "not_mentioned"},"type": "assembly"}},"delivery": {"DB_request": {"req": {"area": "","location": ""},"opt": {"sender": "","recipient": ""}},"T_inform": {"req": {"object": ""},"opt": {"color": "","size": ""},"type": ""}}},"search_result": {"producttype": "detected","product": "detected","quantity": "detected"}},{"user": "All right. Please do that now.","system": "Sure. I'm ready.","s_system": "What a wonderful day!","slots": {"assembly": {"DB_request": {"req": {"producttype": "phone"},"opt": {}},"T_inform": {"req": {"product": "phone","quantity": "five"},"opt": {"color": "black","holeonPCB": "not_mentioned","deadline": "not_mentioned"},"type": "assembly"}},"delivery": {"DB_request": {"req": {"area": "","location": ""},"opt": {"sender": "","recipient": ""}},"T_inform": {"req": {"object": ""},"opt": {"color": "","size": ""},"type": ""}}},"search_result": {"producttype": "detected","product": "detected","quantity": "detected"}},{"user": "I really appreciate your help.","system": "It was my pleasure. I enjoyed helping you.","s_system": "Good bye.","slots": {"assembly": {"DB_request": {"req": {"producttype": "phone"},"opt": {}},"T_inform": {"req": {"product": "phone","quantity": "five"},"opt": {"color": "black","holeonPCB": "not_mentioned","deadline": "not_mentioned"},"type": "assembly"}},"delivery": {"DB_request": {"req": {"area": "","location": ""},"opt": {"sender": "","recipient": ""}},"T_inform": {"req": {"object": ""},"opt": {"color": "","size": ""},"type": ""}}},"search_result": {"producttype": "detected","product": "detected","quantity": "detected"}}]}

FIGURE 9. Dialogue examples of delivery and assembly.

Those metrics are frequently employed to assess the performance of dialogue modules on specific datasets. The first two metrics are designed to assess the accuracy with which the dialogue belief states are captured (see task 1 of figure 8), as well as the fluency and naturalness of the generated response (see task 3 and 4 of figure 8) and exponentiation of the cross-entropy between the word and language model predictions, are assessed by BLEU and perplexity, respectively.

We trained the three versions of GPT neural architectures, gpt2, gpt2-medium, and gpt2-large. We conduct experiments to fine-tune GPT models by varying the percentage of training samples, ranging from 30% to 100%, on IRWoZ dataset (see Table 6). The objective is to evaluate whether the size of IRWoZ dataset will have a significant impact on the performance of the GPT models.

C. RESULTS AND DISCUSSION

As we can see from the results presented in Table 3 and Table 4, the gpt2 model outperforms the gpt2-medium and gpt2-large models on perplexity when trained on 30% and 50% of IRWoZ dataset, respectively. However, the gpt2 model reaches the lowest BLEU scores compared to the other two models. The gpt2-large model has the best overall BLEU scores, while the gpt2-medium model has the highest joint goal accuracy and slot accuracy scores.

When we looked at the results of the Table 5 and Table 6, we discovered that when trained on the full IRWoZ dataset, all models get the same perplexity score. Besides, it is noteworthy that the gpt2-large model does not have a significant impact on perplexity when trained on the full IRWoZ dataset.

Domain	Raw Dialogue Samples
Relocation	"10-Nov-2021-19-10-15.json": {"domain": {"assembly": false, "delivery": false, "position": false, "relocation": true}, "turn": [{"user": "hey robot, can you give that thing?", "system": "sorry, can you describe it a bit?", "s_system": "", "slots": {"assembly": {"DB_request": {"req": {"producttype": ""}, "opt": {}}, "T_inform": {"req": {"quantity": ""}, "opt": {"color": "", "style": "", "size": ""}, "type": ""}}, "delivery": {"DB_request": {"req": {"area": "", "location": ""}, "opt": {"sender": "", "recipient": ""}}, "T_inform": {"req": {"object": ""}, "opt": {"color": "", "size": ""}, "type": ""}}, "position": {"DB_request": {"req": {"position_name": ""}, "opt": {}}, "T_inform": {"req": {"operation": ""}, "opt": {}, "type": ""}}, "relocation": {"DB_request": {"req": {"object_name": "not_mentioned"}, "opt": {}}, "T_inform": {"req": {}, "opt": {"color": "not_mentioned", "size": "not_mentioned", "from": "not_mentioned", "to": "not_mentioned"}, "type": "relocation"}}}, "search_result": {"object_name": "null"}}, {"user": "ah, yes, it is a mouse", "system": "Yes. ", "s_system": "That's I knew.", "slots": {"assembly": {"DB_request": {"req": {"producttype": ""}, "opt": {}}, "T_inform": {"req": {"quantity": ""}, "opt": {"color": "", "style": "", "size": ""}, "type": ""}}, "delivery": {"DB_request": {"req": {"area": "", "location": ""}, "opt": {"sender": "", "recipient": ""}}, "T_inform": {"req": {"object": ""}, "opt": {"color": "", "size": ""}, "type": ""}}, "position": {"DB_request": {"req": {"position_name": ""}, "opt": {}}, "T_inform": {"req": {"operation": ""}, "opt": {}, "type": ""}}, "relocation": {"DB_request": {"req": {"object_name": "mouse"}, "opt": {}}, "T_inform": {"req": {}, "opt": {"color": "not_mentioned", "size": "not_mentioned", "from": "not_mentioned", "to": "not_mentioned"}, "type": "relocation"}}}, "search_result": {"object_name": "detected"}}, {"user": "thank you for your help", "system": "my pleasure ", "s_system": "I am happy to help", "slots": {"assembly": {"DB_request": {"req": {"producttype": ""}, "opt": {}}, "T_inform": {"req": {"quantity": ""}, "opt": {"color": "", "style": "", "size": ""}, "type": ""}}, "delivery": {"DB_request": {"req": {"area": "", "location": ""}, "opt": {"sender": "", "recipient": ""}}, "T_inform": {"req": {"object": ""}, "opt": {"color": "", "size": ""}, "type": ""}}, "position": {"DB_request": {"req": {"position_name": ""}, "opt": {}}, "T_inform": {"req": {"operation": ""}, "opt": {}, "type": ""}}, "relocation": {"DB_request": {"req": {"object_name": "mouse"}, "opt": {}}, "T_inform": {"req": {}, "opt": {"color": "not_mentioned", "size": "not_mentioned", "from": "not_mentioned", "to": "not_mentioned"}, "type": "relocation"}}}, "search_result": {"object_name": "detected"}}]}
Position	"17-Nov-2021-10-56-16.json": {"domain": {"assembly": false, "delivery": false, "position": true, "relocation": false}, "turn": [{"user": "Hey robot, Casper is waiting for you at robot cage.", "system": "Sure", "s_system": "It is a quite busy area.", "slots": {"assembly": {"DB_request": {"req": {"producttype": ""}, "opt": {}}, "T_inform": {"req": {"quantity": ""}, "opt": {"color": "", "style": "", "size": ""}, "type": ""}}, "delivery": {"DB_request": {"req": {"area": "", "location": ""}, "opt": {"sender": "", "recipient": ""}}, "T_inform": {"req": {"object": ""}, "opt": {"color": "", "size": ""}, "type": ""}}, "position": {"DB_request": {"req": {"position_name": "robot cage"}, "opt": {}}, "T_inform": {"req": {"operation": "move"}, "opt": {}, "type": "position"}}, "relocation": {"DB_request": {"req": {"object": ""}, "opt": {}}, "T_inform": {"req": {}, "opt": {"color": "", "size": "", "from": "", "to": ""}, "type": ""}}}, "search_result": {"position_name": "detected", "operation": "detected"}}, {"user": "thanks.", "system": "you are welcome.", "s_system": "looks like it is going to be a busy day.", "slots": {"assembly": {"DB_request": {"req": {"producttype": ""}, "opt": {}}, "T_inform": {"req": {"quantity": ""}, "opt": {"color": "", "style": "", "size": ""}, "type": ""}}, "delivery": {"DB_request": {"req": {"area": "", "location": ""}, "opt": {"sender": "", "recipient": ""}}, "T_inform": {"req": {"object": ""}, "opt": {"color": "", "size": ""}, "type": ""}}, "position": {"DB_request": {"req": {"position_name": "robot cage"}, "opt": {}}, "T_inform": {"req": {"operation": "move"}, "opt": {}, "type": "position"}}, "relocation": {"DB_request": {"req": {"object": ""}, "opt": {}}, "T_inform": {"req": {}, "opt": {"color": "", "size": "", "from": "", "to": ""}, "type": ""}}}, "search_result": {"position_name": "detected", "operation": "detected"}}]}

FIGURE 10. Dialogue examples of relocation and position.

However, it outperformed the other two models on BLEU scores, joint goal accuracy, and slot accuracy for 70% and full IRWoZ dataset, respectively.

From the above results, we noticed a consistent drop in perplexity and a rise in BLEU for all models as a result of increasing the size of the training dataset, especially for gpt2-large. It shows that a larger language model has a greater capacity to learn from a limited dataset and has a more robust generalization capability with fewer data samples. Furthermore, the overall joint goal accuracy and slot accuracy are in a relative high level among all the results, it shows that the two-trail annotation and validation processes contributing to the collection of cleaner data that allows for the better model performance. This encourages further works with focus on dataset's expansion, filter inappropriate responses and correct mislabeled dialogues.

VII. CONCLUSION

The lack of dialogue corpus for training dialogue systems is one of the issues associated with developing dialogue systems in HRI for industrial robots. To cope with this issue, we investigate the capabilities of industrial robots and identify four appropriate scenarios where the dialogue systems can help achieving an easier and more flexible interaction with industrial robots. A human-to-human approach is leveraged with the assistance of the developed IRWoZ web application to facilitate the dialogue corpus collection, and annotation processes. Two-phase trail is conducted with aim of obtaining clean and less noisy dialogues. Additionally, we introduce human conversation strategies and small talk principles to response generation process to obtain a more natural and humanized the dialogue corpus.

<table>
<tr><th>Domain</th><th colspan="2">Annotated Dialogue Samples</th></tr>
<tr><td rowspan="4">Delivery</td><td>Turn 1</td><td><|endoftext|> <|boc|> <|user|> could you deliver the box for me please ? <|eoc|> <|bob|> <|DB_req|> delivery area=not_mentioned location=not_mentioned <|T_req|> delivery object=box <|eob|> <|bosys_act|> delivery area=null location=null object=null <|eosys_act|> <|boTres|> sure , where do you want to deliver it ? <|eoTres|> <|boSres|> <|eoSres|> <|endoftext|></td></tr>
<tr><td>Turn 2</td><td><|endoftext|> <|boc|> <|user|> could you deliver the box for me please ? <|sys|> sure , where do you want to deliver it ? <|user|> yes , it is office in the lab . <|eoc|> <|bob|> <|DB_req|> delivery area=lab location=office <|T_req|> delivery object=box <|eob|> <|bosys_act|> delivery area=detected location=detected object=null <|eosys_act|> <|boTres|> ok , i found location [location] in the [area] . <|eoTres|> <|boSres|> i am so happy! <|eoSres|> <|endoftext|></td></tr>
<tr><td>Turn 3</td><td><|endoftext|> <|boc|> <|user|> could you deliver the box for me please ? <|sys|> sure , where do you want to deliver it ? <|user|> yes , it is office in the lab . <|sys|> ok , i found location office in the lab . i am so happy! <|user|> great , please delivery it now <|eoc|> <|bob|> <|DB_req|> delivery area=lab location=office <|T_req|> delivery object=box <|eob|> <|bosys_act|> delivery area=detected location=detected object=null <|eosys_act|> <|boTres|> sure , i can execute the mission , but i need to know the recipient . <|eoTres|> <|boSres|> sorry for time <|eoSres|> <|endoftext|></td></tr>
<tr><td>Turn 4</td><td><|endoftext|> <|boc|> <|user|> could you deliver the box for me please ? <|sys|> sure , where do you want to deliver it ? <|user|> yes , it is office in the lab . <|sys|> ok , i found location office in the lab . i am so happy! <|user|> great , please delivery it now <|sys|> sure , i can execute the mission , but i need to know the recipient . sorry for time <|user|> yes , send it to tom please . <|eoc|> <|bob|> <|DB_req|> delivery area=lab location=office <|DB_opt|> delivery recipient=Tom <|T_req|> delivery object=box <|eob|> <|bosys_act|> delivery area=detected location=detected recipient=detected object=null <|eosys_act|> <|boTres|> sure , can i do the mission now ? <|eoTres|> <|boSres|> if you would <|eoSres|> <|endoftext|></td></tr>
<tr><td rowspan="3">Assembly</td><td>Turn 1</td><td><|endoftext|> <|boc|> <|user|> how are you , friend ! can you assembly 5 black phones for me ? <|eoc|> <|bob|> <|DB_req|> assembly producttype=not_mentioned <|T_req|> assembly product=phones quantity=five <|T_opt|> assembly color=black <|eob|> <|bosys_act|> assembly producttype=detected product=detected quantity=detected <|eosys_act|> <|boTres|> of course . when do you want me to do ? <|eoTres|> <|boSres|> i am free now . <|eoSres|> <|endoftext|></td></tr>
<tr><td>Turn 2</td><td><|endoftext|> <|boc|> <|user|> how are you , friend ! can you assembly 5 black phones for me ? <|sys|> of course . when do you want me to do ? i am free now . <|user|> all right . please do that now . <|eoc|> <|bob|> <|DB_req|> assembly producttype=phone <|T_req|> assembly product=phone quantity=five <|T_opt|> assembly color=black <|eob|> <|bosys_act|> assembly producttype=detected product=detected quantity=detected <|eosys_act|> <|boTres|> sure . i am ready . <|eoTres|> <|boSres|> what a wonderful day ! <|eoSres|> <|endoftext|></td></tr>
<tr><td>Turn 3</td><td><|endoftext|> <|boc|> <|user|> how are you , friend ! can you assembly 5 black phones for me ? <|sys|> of course . when do you want me to do ? i am free now . <|user|> all right . please do that now . <|sys|> sure . i am ready . what a wonderful day ! <|user|> i really appreciate your help . <|eoc|> <|bob|> <|DB_req|> assembly producttype=phone <|T_req|> assembly product=phone quantity=five <|T_opt|> assembly color=black <|eob|> <|bosys_act|> assembly producttype=detected product=detected quantity=detected <|eosys_act|> <|boTres|> it was my pleasure . i enjoyed helping you . <|eoTres|> <|boSres|> goodbye . <|eoSres|> <|endoftext|></td></tr>
<tr><td rowspan="3">Relocation</td><td>Turn 1</td><td><|endoftext|> <|boc|> <|user|> hey robot , can you give that thing ? <|eoc|> <|bob|> <|DB_req|> relocation object_name=not_mentioned <|T_req|> relocation <|eob|> <|bosys_act|> relocation object_name=null <|eosys_act|> <|boTres|> sorry , can you describe it a bit ? <|eoTres|> <|boSres|> <|eoSres|> <|endoftext|></td></tr>
<tr><td>Turn 2</td><td><|endoftext|> <|boc|> <|user|> hey robot , can you give that thing ? <|sys|> sorry , can you describe it a bit ? <|user|> ah , yes , it is a mouse <|eoc|> <|bob|> <|DB_req|> relocation object_name=mouse <|T_req|> relocation <|eob|> <|bosys_act|> relocation object_name=detected <|eosys_act|> <|boTres|> yes . <|eoTres|> <|boSres|> that s i knew . <|eoSres|> <|endoftext|></td></tr>
<tr><td>Turn 3</td><td><|endoftext|> <|boc|> <|user|> hey robot , can you give that thing ? <|sys|> sorry , can you describe it a bit ? <|user|> ah , yes , it is a mouse <|sys|> yes . that s i knew . <|user|> thank you for your help <|eoc|> <|bob|> <|DB_req|> relocation object_name=mouse <|T_req|> relocation <|eob|> <|bosys_act|> relocation object_name=detected <|eosys_act|> <|boTres|> my pleasure <|eoTres|> <|boSres|> i am happy to help <|eoSres|> <|endoftext|></td></tr>
<tr><td rowspan="2">Position</td><td>Turn 1</td><td><|endoftext|> <|boc|> <|user|> hey robot , casper is waiting for you at robot cage . <|eoc|> <|bob|> <|DB_req|> position position_name=robot cage <|T_req|> position operation=move <|eob|> <|bosys_act|> position position_name=detected operation=detected <|eosys_act|> <|boTres|> sure <|eoTres|> <|boSres|> it is a quite busy area . <|eoSres|> <|endoftext|></td></tr>
<tr><td>Turn 2</td><td><|endoftext|> <|boc|> <|user|> hey robot , casper is waiting for you at robot cage . <|sys|> sure it is a quite busy area . <|user|> thanks . <|eoc|> <|bob|> <|DB_req|> position position_name=robot cage <|T_req|> position operation=move <|eob|> <|bosys_act|> position position_name=detected operation=detected <|eosys_act|> <|boTres|> you are welcome . <|eoTres|> <|boSres|> looks like it is going to be a busy day . <|eoSres|> <|endoftext|></td></tr>
</table>

FIGURE 11. **Dialogue examples marked with special tokens.**

We run three experiments with the SOTA GPT language models on the IRWoZ dataset. The evaluation results indicate that the models can reach high accuracy on our dataset, at the same time being efficient in terms of fluency and natural response generation. Our work marks an important step towards supporting the speech-enabled dialogue systems for

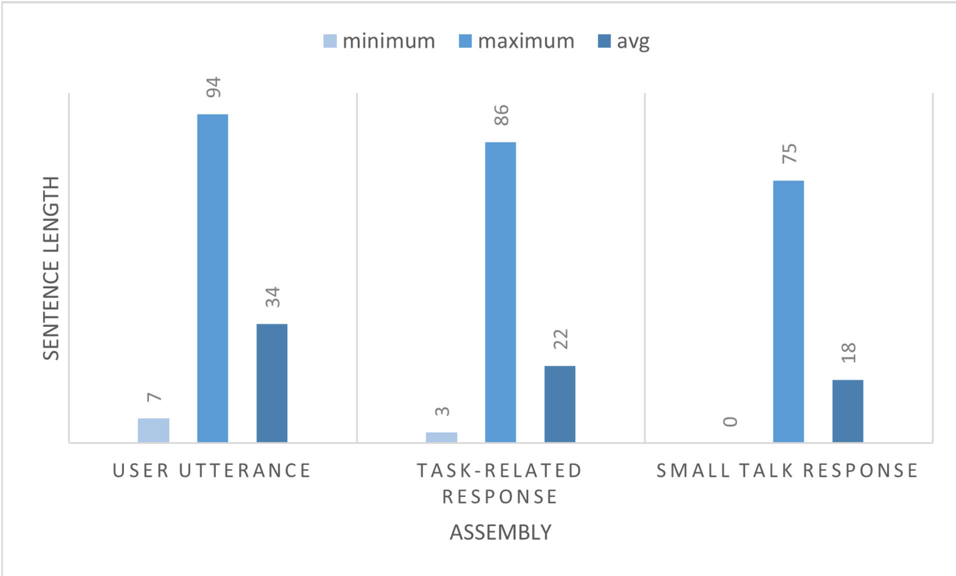


FIGURE 12. **Task assembly: distribution of sentence length.**

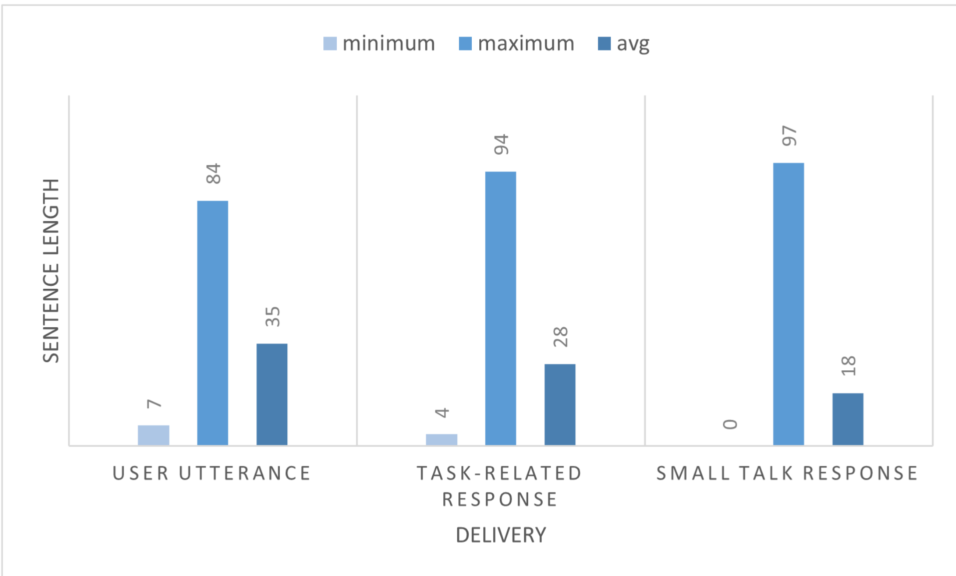


FIGURE 13. **Task delivery: distribution of sentence length.**

manufacturing, particularly with regard to HRI in industrial robots. We provide the proposed dataset as open-source to the public, to serve as a new benchmark for better model and performance comparisons in design and development of industrial robot dialogue systems as well as other dialogue downstream tasks.

Though the IRWoZ is the first dialogue corpus proposed for industrial purposes, the size of the dialogues is limited, as well as the domains. Therefore, our future work would aim to expand and improve the current dataset by collecting more high-quality data, and to explore new areas. First, We are going to expand the current dataset by organizing a large-scale WoZ simulation for data collection and examining additional areas such as navigation and inspection. Second, we would need to analyze the collected data to identify patterns and trends, and to evaluate the effectiveness of the WoZ approach for data collection in the context of navigation and inspection. Third, we could then refine and improve the WoZ approach for data collection in this context, by developing new techniques or modifying existing ones. This could involve developing new task scenarios, improving the dialogue system, or refining the interaction protocol between the human Wizard and the participants. This would have important implications for the development of dialogue

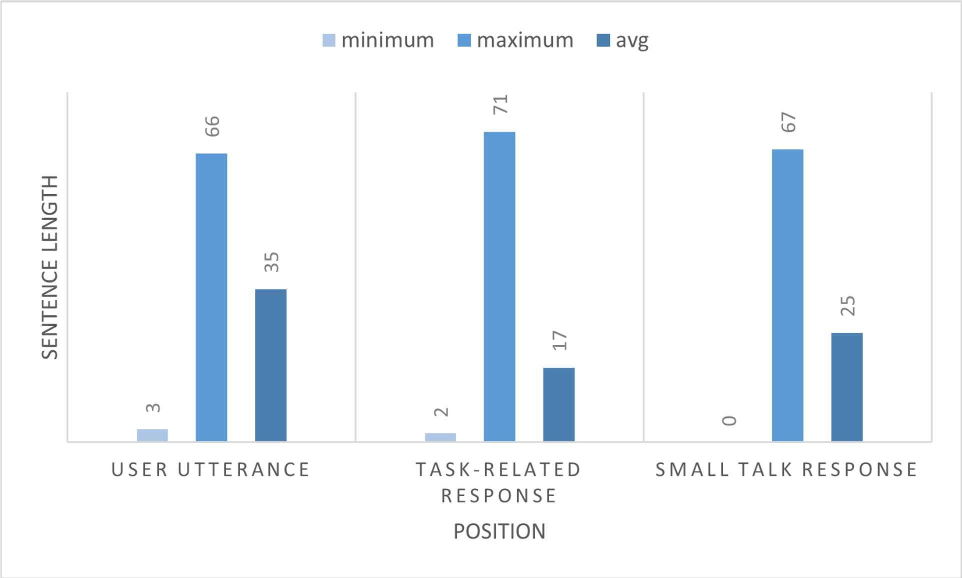


FIGURE 14. Task position: distribution of sentence length.

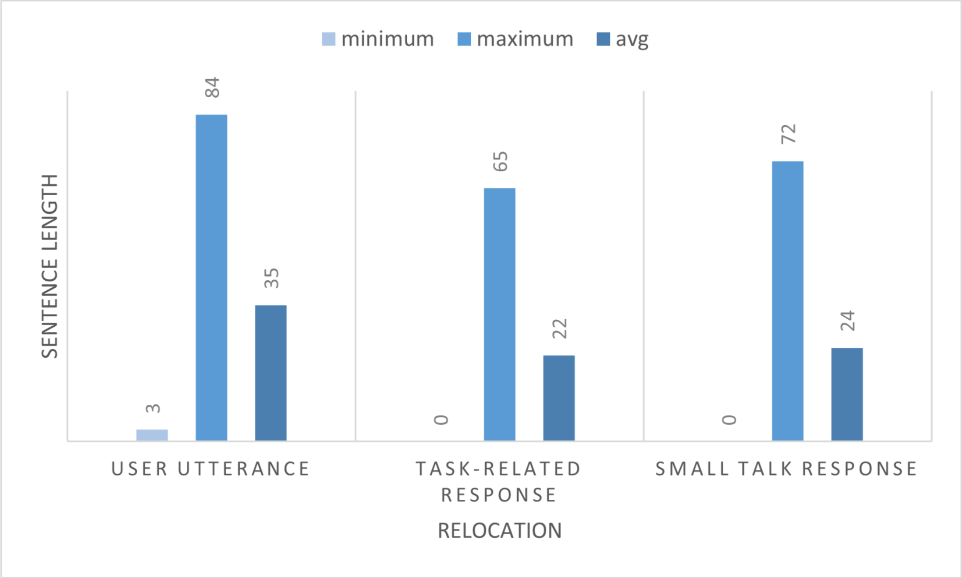


FIGURE 15. Task relocation: distribution of sentence length.

systems in these domains, and could lead to more effective and efficient systems that are better able to meet user needs and preferences.

APPENDIX A
RAW DIALOGUE SAMPLES

Figure 9 and Figure 10 show the raw dialogue examples of each domain, respectively.

APPENDIX B
ANNOTATED DIALOGUE SAMPLES

Figure 11 shows the marked dialogue examples of appendix A with special tokens.

APPENDIX C
DISTRIBUTION OF SENTENCE LENGTH

See Figures 12–15.

REFERENCES

[1] O. Madsen, C. Bro Sørensen, R. Larsen, L. Overgaard, and N. J. Jacobsen, "A system for complex robotic welding," *Ind. Robot, Int. J.*, vol. 29, no. 2, pp. 127–131, Apr. 2002.

[2] A. S. Olesen, B. B. Gergaly, E. A. Ryberg, M. R. Thomsen, and D. Chrysostomou, "A collaborative robot cell for random bin-picking based on deep learning policies and a multi-gripper switching strategy," *Proc. Manuf.*, vol. 51, pp. 3–10, Jan. 2020.

[3] J. F. Buhl, R. Grønhøj, J. K. Jørgensen, G. Mateus, D. Pinto, J. K. Sørensen, S. Bøgh, and D. Chrysostomou, "A dual-arm collaborative robot system for the smart factories of the future," *Proc. Manuf.*, vol. 38, pp. 333–340, Jan. 2019.

[4] E. R. da Silva, C. Schou, S. Hjorth, F. Tryggvason, and M. S. Sørensen, "Plug & produce robot assistants as shared resources: A simulation approach," *J. Manuf. Syst.*, vol. 63, pp. 107–117, Apr. 2022.

[5] S. Bogh, "Integration and assessment of multiple mobile manipulators in a real-world industrial production facility," in *Proc. ISR/Robotik 41st Int. Symp. Robot.*, Jun. 2014, pp. 1–8.

[6] S. Hjorth, J. Lachner, S. Stramigioli, O. Madsen, and D. Chrysostomou, "An energy-based approach for the integration of collaborative redundant robots in restricted work environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 7152–7158.

[7] C. Li and H. J. Yang, "Bot-X: An AI-based virtual assistant for intelligent manufacturing," *Multiagent Grid Syst.*, vol. 17, no. 1, pp. 1–14, Apr. 2021.

[8] M. Dibitonto, K. Leszczynska, F. Tazzi, and C. M. Medaglia, "Chatbot in a campus environment: Design of LiSA, a virtual assistant to help students in their university life," in *Proc. Int. Conf. Hum.-Comput. Interact.* Cham, Switzerland: Springer, 2018, pp. 103–116.

[9] G. Iannizzotto, L. L. Bello, A. Nucita, and G. M. Grasso, "A vision and speech enabled, customizable, virtual assistant for smart environments," in *Proc. 11th Int. Conf. Hum. Syst. Interact. (HSI)*, Jul. 2018, pp. 50–56.

[10] M. Duguleană, V.-A. Briciu, I.-A. Duduman, and O. M. Machidon, "A virtual assistant for natural interactions in museums," *Sustainability*, vol. 12, no. 17, p. 6958, Aug. 2020.

[11] K. Laeeq and Z. A. Memon, "Scavenge: An intelligent multi-agent based voice-enabled virtual assistant for LMS," *Interact. Learn. Environ.*, vol. 29, no. 6, pp. 954–972, Aug. 2021.

[12] C. Li, J. Park, H. Kim, and D. Chrysostomou, "How can I help you? An intelligent virtual assistant for industrial robots," in *Proc. Companion ACM/IEEE Int. Conf. Hum.-Robot Interact.*, Mar. 2021, pp. 220–224, doi: 10.1145/3434074.3447163.

[13] D. Evangelista, W. Villa, M. Imperoli, A. Vanzo, L. Iocchi, D. Nardi, and A. Pretto, "Grounding natural language instructions in industrial robotics," in *Proc. IEEE/RSJ IROS Workshop, Hum.-Robot Interact. Collaborative Manuf. Environ.*, 2017, pp. 1–6.

[14] Y. Lin, H. Zhou, M. Chen, and H. Min, "Automatic sorting system for industrial robot with 3D visual perception and natural language interaction," *Meas. Control*, vol. 52, nos. 1–2, pp. 100–115, Jan. 2019.

[15] C. Li, A. K. Hansen, D. Chrysostomou, S. Bogh, and O. Madsen, "Bringing a natural language-enabled virtual assistant to industrial mobile robots for learning, training and assistance of manufacturing tasks," in *Proc. IEEE/SICE Int. Symp. Syst. Integr. (SII)*, Jan. 2022, pp. 238–243.

[16] S. Lu, J. Berger, and J. Schilp, "System of robot learning from multi-modal demonstration and natural language instruction," *Proc. CIRP*, vol. 107, pp. 914–919, Jan. 2022.

[17] Y. Lin, H. Min, H. Zhou, and M. Chen, "A natural language interaction based automatic operating system for industrial robot," in *Proc. Int. Conf. Intell. Comput.* Cham, Switzerland: Springer, 2018, pp. 111–122.

[18] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gašić, "MultiWOZ—A large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 1–11.

[19] E. Hosseini-Asl, B. McCann, C. S. Wu, S. Yavuz, and R. Socher, "A simple language model for task-oriented dialogue," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2020, pp. 20179–20191.

[20] B. Peng, C. Li, J. Li, S. Shayandeh, L. Liden, and J. Gao, "SOLOIST: Building task bots at scale with transfer learning and machine teaching," 2020, *arXiv:2005.05298*.

[21] I. V. Serban, R. Lowe, P. Henderson, L. Charlin, and J. Pineau, "A survey of available corpora for building data-driven dialogue systems," 2015, *arXiv:1512.05742*.

[22] H. Chen, X. Liu, D. Yin, and J. Tang, "A survey on dialogue systems: Recent advances and new frontiers," *ACM SIGKDD Explor. Newslett.*, vol. 19, no. 2, pp. 25–35, Dec. 2017.

[23] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, "The ATIS spoken language systems pilot corpus," in *Proc. Workshop Speech Natural Lang. (HLT)*, Jun. 1990, pp. 1–6. [Online]. Available: https://aclanthology.org/H90-1021

[24] M. Henderson, B. Thomson, and J. D. Williams, "The second dialog state tracking challenge," in *Proc. 15th Annu. Meeting Special Interest Group Discourse Dialogue (SIGDIAL)*, 2014, pp. 263–272.

[25] T.-H. Wen, D. Vandyke, N. Mrksic, M. Gasic, L. M. Rojas-Barahona, P.-H. Su, S. Ultes, and S. Young, "A network-based end-to-end trainable task-oriented dialogue system," 2016, *arXiv:1604.04562*.

[26] M. Eric and C. D. Manning, "Key-value retrieval networks for task-oriented dialogue," 2017, *arXiv:1705.05414*.

[27] H. Zhou, C. Zheng, K. Huang, M. Huang, and X. Zhu, "KdConv: A Chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation," 2020, *arXiv:2004.04100*.

[28] X. Li, Y.-N. Chen, L. Li, J. Gao, and A. Celikyilmaz, "End-to-end task-completion neural dialogue systems," in *Proc. 8th Int. Joint Conf. Natural Lang. Process.*, 2017, pp. 1–7.

[29] C.-S. Wu, S. Hoi, R. Socher, and C. Xiong, "TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue," 2020, *arXiv:2004.06871*.

[30] Z. Lin, A. Madotto, G. I. Winata, and P. Fung, "MinTL: Minimalist transfer learning for task-oriented dialogue systems," 2020, *arXiv:2009.12005*.

[31] L. Chen, B. Lv, C. Wang, S. Zhu, B. Tan, and K. Yu, "Schema-guided multi-domain dialogue state tracking with graph attention neural networks," in *Proc. AAAI*, 2020, pp. 1–8.

[32] B. Liu, G. Tür, D. Hakkani-Tür, P. Shah, and L. Heck, "Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., (Long Papers)*, vol. 1, 2018, pp. 2060–2069. [Online]. Available: https://aclanthology.org/N18-1187

[33] B. Dhingra, "Towards end-to-end reinforcement learning of dialogue agents for information access," in *Proc. ACL*, Jul. 2017, pp. 484–495. [Online]. Available: https://aclanthology.org/P17-1045

[34] Q. Wu, Y. Zhang, Y. Li, and Z. Yu, "Alternating recurrent dialog model with large-scale pre-trained language models," 2019, *arXiv:1910.03756*.

[35] M. Richardson, C. J. Burges, and E. Renshaw, "Mctest: A challenge dataset for the open-domain machine comprehension of text," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 193–203.

[36] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," 2017, *arXiv:1710.03957*.

[37] C. Danescu-Niculescu-Mizil, J. Cheng, J. Kleinberg, and L. Lee, "You had me at hello: How phrasing affects memorability," in *Proc. ACL*, 2012, pp. 892–901.

[38] E. N. Forsythand and C. H. Martell, "Lexical and discourse analysis of online chat dialog," in *Proc. Int. Conf. Semantic Comput. (ICSC)*, Sep. 2007, pp. 19–26.

[39] M. Elsner and E. Charniak, "You talking to me? A corpus and algorithm for conversation disentanglement," in *Proc. ACL*, 2008, pp. 1–9.

[40] Y. Meng, S. Wang, Q. Han, X. Sun, F. Wu, R. Yan, and J. Li, "OpenViDial: A large-scale, open-domain dialogue dataset with visual contexts," 2020, *arXiv:2012.15015*.

[41] C.-H. Lee, S.-M. Wang, H.-C. Chang, and H.-Y. Lee, "ODSQA: Open-domain spoken question answering dataset," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2018, pp. 949–956.

[42] V. Logacheva, V. Malykh, A. Litinsky, and M. Burtsev, "ConvAI2 dataset of non-goal-oriented human-to-bot dialogues," in *The NeurIPS'18 Competition*. Cham, Switzerland: Springer, 2020, pp. 277–294.

[43] F. Radlinski, K. Balog, B. Byrne, and K. Krishnamoorthi, "Coached conversational preference elicitation: A case study in understanding movie preferences," in *Proc. 20th Annu. SIGdial Meeting Discourse Dialogue*, 2019, pp. 353–360. [Online]. Available: https://aclanthology.org/W19-5941

[44] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, "Wizard of Wikipedia: Knowledge-powered conversational agents," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–18. [Online]. Available: https://openreview.net/forum?id=r1l73iRqKm

[45] M. R. Pedersen, "Robot skills for manufacturing: From concept to industrial deployment," *Robot. Comput.-Integr. Manuf.*, vol. 37, pp. 282–291, Feb. 2016.

[46] J. O. Huckaby and H. I. Christensen, "A taxonomic framework for task modeling and knowledge transfer in manufacturing robotics," in *Proc. Workshops 26th AAAI Conf. Artif. Intell.*, 2012, pp. 1–8.

[47] M. Pantano, T. Eiband, and D. Lee, "Capability-based frameworks for industrial robot skills: A survey," 2022, *arXiv:2203.00538*.

[48] P. Segura, O. Lobato-Calleros, A. Ramírez-Serrano, and I. Soria, "Human-robot collaborative systems: Structural components for current manufacturing applications," *Adv. Ind. Manuf. Eng.*, vol. 3, Nov. 2021, Art. no. 100060.

[49] Z. Yu, Z. Xu, A. W. Black, and A. Rudnicky, "Strategy and policy learning for non-task-oriented conversational systems," in *Proc. 17th Annu. Meeting Special Interest Group Discourse Dialogue*, 2016, pp. 404–412.

[50] C. Fleming, *It's Way You Say It: Becoming Articulate, Well-Spoken, Clear*. Oakland, CA, USA: Berrett-Koehler Publishers, 2013.

[51] C. Li, X. Zhang, D. Chrysostomou, and H. Yang, "ToD4IR: A humanised task-oriented dialogue system for industrial robots," *IEEE Access*, vol. 10, pp. 91631–91649, 2022.

[52] K. Sun, S. Moon, P. Crook, S. Roller, B. Silvert, B. Liu, Z. Wang, H. Liu, E. Cho, and C. Cardie, "Adding chit-chat to enhance task-oriented dialogues," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2021, pp. 1–14.

[53] J. F. Kelley, "An iterative design methodology for user-friendly natural language office information applications," *ACM Trans. Inf. Syst.*, vol. 2, no. 1, pp. 26–41, Jan. 1984, doi: 10.1145/357417.357420.

[54] M. Eric, R. Goel, S. Paul, A. Kumar, A. Sethi, P. Ku, A. K. Goyal, S. Agarwal, S. Gao, and D. Hakkani-Tur, "MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines," 2019, *arXiv:1907.01669*.

[55] X. Zang, A. Rastogi, and J. Chen, "MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines," in *Proc. 2nd Workshop Natural Lang. Process. Conversational AI*, 2020, pp. 109–117.

[56] A. Radford, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[57] T.-R. Chiang and Y.-T. Yeh, "Improving dialogue state tracking by joint slot modeling," 2021, *arXiv:2109.14144*.

[58] L. Reed, S. Oraby, and M. Walker, "Can neural generators for dialogue learn sentence planning and discourse structuring?" 2018, *arXiv:1809.03015*.

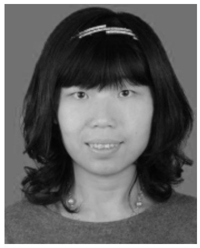
[59] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2001, pp. 311–318.

[60] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

CHEN LI (Member, IEEE) received the M.S. degree in computer application from the University of Shanghai for Science and Technology, Shanghai, China, in 2010, and the Ph.D. degree in computer science and technology from Shanghai Jiao Tong University, Shanghai, in 2015.

From 2015 to 2016, he was a Research Assistant with the Centre for Creative Computing, Bath Spa University. From 2016 to 2018, he was a Research Associate with the Department of Computer Science, Imperial College London. From 2018 to 2020, he held a postdoctoral position with the Department of Materials and Production, Aalborg University, where he has been an Assistant Professor, since 2020. His research interests include natural language processing, human–robot interaction, and system modeling.

DIMITRIOS CHRYSOSTOMOU (Member, IEEE) received the Diploma degree in production engineering, in 2006, and the Ph.D. degree in robot vision from the Democritus University of Thrace, Greece, in 2013. Since 2013, he has been with the Robotics and Automation Group, Department of Materials and Production, Aalborg University, Denmark, where he was a Postdoctoral Researcher, from 2013 to 2016; an Assistant Professor, from 2016 to 2019; an Associate Professor, and an Associate Professor, since 2020. His research interests include robot vision, skill-based programming, and human–robot interaction for intelligent robot assistants.



XIAOCHUN ZHANG (Member, IEEE) received the Ph.D. degree from the School of Computer Science and Technology, Nanjing University of Science and Technology, in 2014. She is currently with the School of Management Science and Computer, Anhui University of Finance and Economics, China. Her research interests include computer vision, time series prediction, and dialogue systems.

HONGJI YANG received the B.S. and M.S. degrees in computer science from Jilin University, Changchun, China, in 1982 and 1985, respectively, and the Ph.D. degree in computer science from Durham University, Durham, U.K., in 1994. He is currently with the School of Computing and Mathematical Sciences, University of Leicester. He has published more than 500 articles. His main research interests include knowledge modeling and creative computing. He became a Golden Core Member of the IEEE Computer Society, in 2010.