IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 August 2022, accepted 17 August 2022, date of publication 29 August 2022, date of current version 6 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3202554

RESEARCH ARTICLE

ToD4IR: A Humanised Task-Oriented Dialogue System for Industrial Robots

CHEN LI[1], (Member, IEEE), XIAOCHUN ZHANG[2], (Member, IEEE), DIMITRIOS CHRYSOSTOMOU[1], (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, Leicester LE1 7RH, U.K.

Corresponding author: Xiaochun Zhang (xiaochun.zhang@aufe.edu.cn)

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

ABSTRACT Despite the fact that task-oriented conversation systems have received much attention from the dialogue research community, only a handful of them have been studied in a real-world manufacturing context using industrial robots. One stumbling block is the lack of a domain-specific discourse corpus for training these systems. Another difficulty is that earlier attempts to integrate natural language interfaces (such as chatbots) into the industrial sector have primarily focused on task completion rates. When designing a dialogue system for social robots, the user experience is prioritized above industrial robots. To overcome these challenges, we provide the Industrial Robots Domain Wizard-of-Oz dataset (IRWoZ), a fully-labeled discourse dataset covering four robotics domains. It delivers simulated discussions between shop floor workers and industrial robots, with over 401 dialogues, to promote language-assisted Human-Robot Interaction (HRI) in industrial settings. Small talk concepts and human-to-human conversation strategies are provided to support human-like answer generation and give a more natural and adaptable dialogue environment to increase user experience and engagement. Finally, we propose and evaluate an end-to-end Task-oriented Dialogue for Industrial Robots (ToD4IR) using two types of pre-trained backbone models: GPT-2 and GPT-Neo, on the IRWoZ dataset. We performed a series of trials to validate ToD4IR's performance in a real manufacturing context. Our experiments demonstrate that ToD4IR outperforms three downstream task-oriented dialogue tasks, i.e., dialogue state tracking, dialogue act generation, and response generation, on the IRWoZ dataset. Our source code of ToD4IR and the IRWoZ dataset is accessible at https://github.com/lcroy/ToD4IR for reproducible research.

INDEX TERMS Natural language processing, interactive systems, data collection, neural networks, human–robot interaction.

I. INTRODUCTION

Conversational artificial intelligence (AI), such as dialogue systems and chatbots, has received much attention in recent years and reaped various positive results [1], [2], [3]. Despite its enormous success, developing robots with human-like natural language response capabilities through conversational AI is still challenging. Many studies have been conducted in four directions to construct human-like dialogue systems that improve the user experience: knowledge, empathy, engagingness, and humanness [4], [5], [6], [7], [8]. The majority of these studies, however, are focused on open-domain dialogue systems, whereas task-oriented dialogue (ToD) aims to meet the functional requirements [3], [9] of specialized domains.

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

Furthermore, some research works have been done on designing and developing such ToD for the manufacturing domain, especially in human-robot interaction (HRI) for industrial robots [10], [11]. While several studies have been presented to ground natural language commands for industrial robot manipulation [12], [13], the majority do not need dialogue datasets to train a neural network since they employ the keywords-matching approach, or their dialogue datasets are not publicly available. Additionally, while adopting language-enabled ToD to improve HRI for industrial jobs is a novel concept, there is currently no pre-built dialogue corpus for training such ToD.

The contributions of our work can be summarized as follows:

- As a new benchmark, we provide the Industrial Robots Wizard-of-Oz dataset (IRWoZ), a scalable, innovative manufacturing domain-centered ToD dataset that may be used to create ToD in HRI for industrial robots. IRWoZ is open to the public, facilitating successful collaboration between academics and industrial partners.
- We introduce small talk principles and human-to-human conversation strategies to assist the human-like response generation build the IRWoZ. To the best of our knowledge, it is the first effort to produce a human-like response in HRI for industrial robots.
- We offer ToD4IR, a conversational AI, fine-tuned on our IRWoZ corpus utilizing two types of State-of-The-Art (SoTA) pre-trained backbone models (GPT-2 and GPT-Neo).
- We demonstrate the robustness of our approach in a real-world factory setup.

The remainder of this work is structured as follows: Section II summarizes the related work. Section III then outlines how we created the IRWoZ dataset, the ToD4IR system architecture, and essential components. Section IV summarizes the findings of the evaluation and discussion. The paper is concluded in Section V.

II. RELATED WORK

A. TASK-ORIENTED DIALOGUE SYSTEMS

In general, ToD aims to accomplish specific tasks through dialogue with the user, such as booking a taxi or ordering food. Such systems are typically classified into two categories: pipeline and end-to-end [14]. Compared to a pipeline manner, partial or complete end-to-end dialogue systems have received more attention in recent years [15]. Each component is trained end-to-end, reducing sub-model errors' propagation and accumulation. Furthermore, passing user feedback to a model is challenging because each module is interdependent in a pipeline pattern [16]. The evaluation of such end-to-end ToD systems mainly focuses on dialogue state tracking and response generation. Lei *et al.* presented a sequence-to-sequence method incorporating dialogue state tracking and response generation [2]. Li *et al.* proposed an end-to-end neural dialogue system for achieving targets and reaching higher accuracy [17]. Perez *et al.* an end-to-end memory network, a memory-enhanced neural network architecture, for dialog state tracking [18]. The proposed Alternating Roles Dialog Model (ARDM) uses the large pre-trained language model. It does not require belief states or dialog acts from human annotations [19]. Chen *et al.* introduced a graph attention network to extract information from utterances and graphs and leveraged a recurrent graph attention network to control state updating [20]. Wu *et al.* proposed a pre-trained ToD-BERT that models dialogue behavior during pre-training and outperforms downstream tasks, e.g., response selection [21]. Unlike TOD-BERT, Minimalist Transfer Learning (MinTL) used a copy mechanism to inject the previous dialogue states into the new one and improve the end-to-end response generation [22]. In SimpleTOD, Hosseini-Asi *et al.* took the whole ToD as a single sequence prediction problem, leveraged transfer learning from a pre-trained language model based on open-domain, and improved the performance for dialogue model [9]. Soloist used task-grounded pre-training to learn tasks while enjoying low annotation cost for the training dataset and reached a higher task success rate [23]. Human feedback is considered during the training stage in the end-to-end model [24] to improve the system performance. Bhuwan *et al.* presented an end-to-end differentiable KB-Infobot, which improved the robustness of the system and flexibility of question types [25].

The most related works to ours are [9], [23], which leverage the pre-trained language model to build end-to-end ToD systems. However, those works do not distinguish the dialogue belief state but treat them as a whole entity. Our work divides the belief state into four groups: database-related, task-related, required, and optional, based on the source and importance, respectively. It helps to generate system actions and responses with higher efficiency and accuracy.

B. CHIT-CHAT DIALOGUE SYSTEMS

In comparison to task-oriented dialogue systems, chit-chat systems place a higher premium on small talk, sociability, engagement, and humanness.

He *et al.* investigated the effects of various pre-trained fine-tune schemes. They found that some significant language generation methods can be forgotten due to data separation. They proposed a mix-review method based on the data mixture idea and effectively alleviated language skill forgotten problems [26]. Daniel *et al.* employed a sequence-to-sequence model, Meena, which includes two 260 million parameters. Compared to other chatbots, Meena is an open-domain model based on a multi-turn transformer architecture; it can give unique and more reasonable responses. The paper also presented a new Sensibleness and Specificity Average (SSA) index for Manual evaluation [27]. Sun *et al.* integrated Chit-Chat to enhance task-oriented dialogue and achieve the goal of making virtual assistant conversations more engaging and interactive [28]. Roller and his research team think good conversation requires varied skills, including engagement, knowledge, empathy, and personality [4].

Zhang *et al.* developed a DIALOGPT model, trained on a massive real-world Reddit dataset [29]. Moon *et al.* proposed the model with light social greetings annotations for a few chit-chat dataset [30]. Another work suggests randomly sampling utterances from a chit-chat corpus to improve the out-of-domain recovery performance [31]. XiaoIce's design considers intelligence and emotion together, which advances communication, engagement, and social belonging [32]. Shu proposed a Meta-Learning framework (DAML) based on random source domains with disparate label sets and achieves high performance on an unknown target domain [33]. Asma proposed a self-play metric where the dialog system talks to itself. They showed that this metric was similar to the human-rated quality for a dialog model and better than other automated metrics [34]. Hannah *et al.* proposed a new model based on empathetic dialogue generation and created a novel dataset of 25k emotional dialogue [5]. Mazare *et al.* attempted to focus on personal facts to make their chit-chat dialogue model more engagements [35]. Akasaki and Kaji annotate user utterances with chat/non-chat binary labels [3].

While the aforementioned chit-chat dialogue systems are primarily concerned with the open domain or function more like personal assistants, borrowing such features (e.g., small talk, engagement) that enable users to speak naturally in order to complete tasks more efficiently is also critical for developing robust and humanized task-oriented dialogue systems.

C. DIALOGUE DATASETS

A dialogue dataset is critical for building data-driven-based conversational AI. Such datasets may cover many different domain categories. Daily dialogue is a human-written dataset based on communication intention and emotional information and achieves good performance in multi-turn dialogues [36]. Cornell Movie-Dialogs corpus is generated in a social context from movies. This corpus contains a large metadata-rich collection of fictional conversations extracted from a raw movie, including 304,713 utterances in total [37]. Sun *et al.* bridge the two baselines, and compare ACCENTOR-SGD and ACCENTOR-MultiWOZ with original SGD [20] and MultiWOZ [38], [39], [40] datasets [28]. The ConvAI2 is based on large pre-trained transformers [41]. BlendedSkillTalk is a conversation dataset of about 7k entries explicitly designed to exhibit multiple conversation modes, focusing on personality, empathy, and knowledge [42]. Bookscorpora includes some fixtures and the integration of proxy usage, which downloads books from the original smashwords dataset where all books are written in English with at least 20k words. Wolf *et al.* used the BooksCorpus dataset to get the best perplexities and F1 scores [43]. The Ubuntu Dialogue Corpus contains over seven million utterances and 100 million words for multi-turn dialogues. It is based on neural language models with large amounts of unlabeled data [44].

While new datasets have been developed constantly, none of them are specifically targeted toward industrial robots. In our study, we create a novel dialogue corpus focusing on four common topics for HRI in industrial robotics. Additionally, we seek to increase shop floor workers' engagement and user experience by offering a natural, humanized conversation environment. That is, to enable ToD4IR to communicate in natural human language. To do this, ToD4IR must be trained on a corpus of natural human-to-human conversations to understand how to generate a human-like response. While the SoTA work [8] incorporates chit-chat to augment its ToD and utilizes pre-trained generative models to produce free-form chit-chat data, it requires a hybrid classifier to filter candidate datasets and crowd workers to annotate data. Nevertheless, our dataset is enriched by including small talk and human-to-human conversation strategies during the dataset collection process. No extra classification or interaction by human workers is required to validate the candidate response, as we collect the discussion corpus directly using "Wizard-of-OZ," a human-to-human technique.

III. METHOD

The approach for producing the IRWoZ corpus, which mixes small talk and human-to-human conversation strategies with task-oriented dialogue and is used in our pre-training, is described in this section. The proposed ToD4IR's overall system architecture, as well as the backbone models trained on a dialogue history level, are then presented.

A. DATASET CREATION

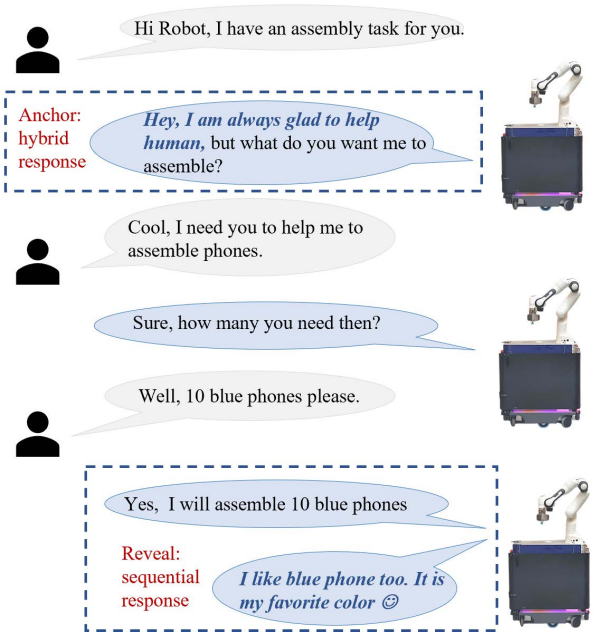
Among our contributions is the development of an industrial robot-oriented dialogue corpus, IRWoZ, a fully-labeled collection of human-to-human conversations spanning over four domains (assembly and relocation tasks of industrial manipulator, delivery and positioning tasks of mobile industrial robots). It seeks to create simulated dialogues between shop floor workers and industrial robots in order to facilitate language-assisted HRI in an industrial setting, with a total of 401 dialogues.

This section describes the four-step process by which the IRWoZ dataset was created. Firstly, we investigated the most typical scenarios for HRI in industrial robots to identify areas ideal for incorporating language-enabled ToD. Secondly, we leverage the WoZ method for collecting domain-specific conversation datasets, which receive less attention and are not readily available online or on the market. Thirdly, we examined work-related small talk principles to guide the humanized response generation to boost the user experience. Finally, we explored human-to-human conversation strategies to provide a more meaningful task-oriented response and maintain a high task completion rate.

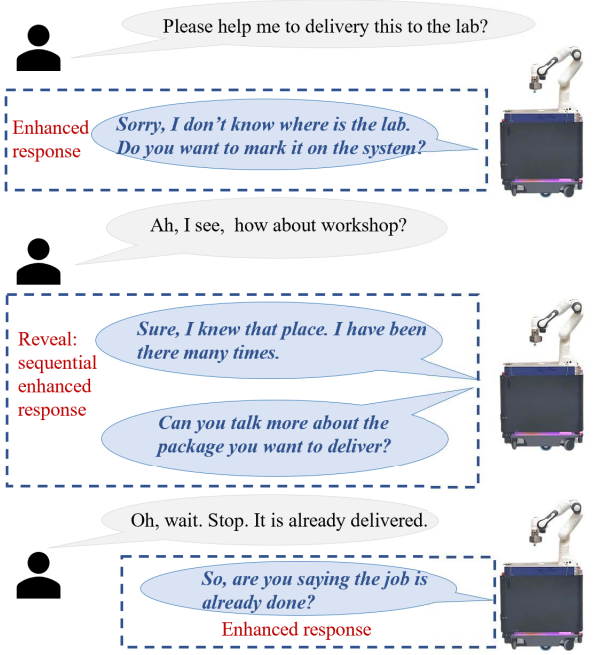
1) SCENARIOS FOR INDUSTRIAL ROBOTS

In general, industrial robots are employed in the following scenarios [45]:

- **Material handling**, including *material transfer*, which requires the robot to move materials or work parts from one location to another, and *machine loading and*



(a) An example dialogue of assembly task. ToD uses the Anchor and Reveal principle to generate the hybrid and sequential responses in turns one and three, respectively.



(b) An example dialogue of delivery task. ToD uses the *End topics with a question*,*Elicit More Information* and *Clarifying* strategies to generate enhanced response respectively.

FIGURE 1. **Two examples of the interactions between a shop floor worker and a ToD4IR system.**

unloading, which utilizes a robot to load and unload parts at a production machine.

- **Processing operations** where a robot manipulates a tool to carry out a process on the work part, e.g., assembly, inspection.
- **Service operations** where a robot fulfills several operational services, e.g., repairing the manufacturing equipment, cleaning up the waste, and scrap after manufacturing tasks.

In comparison to machine loading and unloading, ToD is more involved in the material transfer scenario, e.g., internal transportation, in which the user requests a delivery task by informing a mobile robot (e.g., Mobile Industrial Robot (MIR)[1]) the destination of the transportation, recipient, and optional information of parcels (e.g., size, color). Fig. 1 (b) depicts an example dialogue of such scenario. Human collaborates with an industrial manipulator for assembly tasks (e.g., Universal robots,[2] Franka Emika[3]), on the other hand, is also common in a production environment where ToD fits well. For instance, a shop floor worker may request a Franka Emika robot to assemble a product via dialogue in which the ToD system gathers the information of the product type and amount and controls the robot to accomplish the assembly operation. An example dialogue for such assembly scenario is illustrated in Fig. 1 (a). However, service operations, such as cleaning and repairing, are too task-specific to be generalized for ToD.

[1]https://www.mobile-industrial-robots.com/da/
[2]https://www.universal-robots.com/
[3]https://www.franka.de/

2) DATA COLLECTION - A WOZ WAY

In general, most of the ToD dataset is built based on existing dialogue systems [38]. To our knowledge, the ToD corpus for HRI in industrial robots is unavailable. Inspired by the Wizard-of-Oz framework (WOZ) [46], MultiWOZ datasets [38], [39], [40] and its succeeded validation in [15], [47], we build a dialogue corpus via human-to-human data collection method. To obtain a more reliable and diverse dialogue corpus, participants invited must have background and skills related to Robotics, Automation, or Computer Science, with an emphasis on expertise in using industrial robots. Furthermore, factory workers and engineers are also included, given that their regular tasks involve operating industrial robots. As a result, an 18-member team is created, consisting of three factory engineers, six Ph.D. students, two master's students, and seven professors. All students have a background in Robotics/Automation and have experience in programming and manipulating industrial robots. Among seven professors, four of them work on industrial robots, one focuses on automation, and the other two have a background in computer science.

As seen in Fig. 2, the simulation environment is Aalborg University's Learning Factory [48]. The robot utilized in the simulation is one of our autonomous industrial mobile manipulators, Little Helper (LH), currently in its eighth

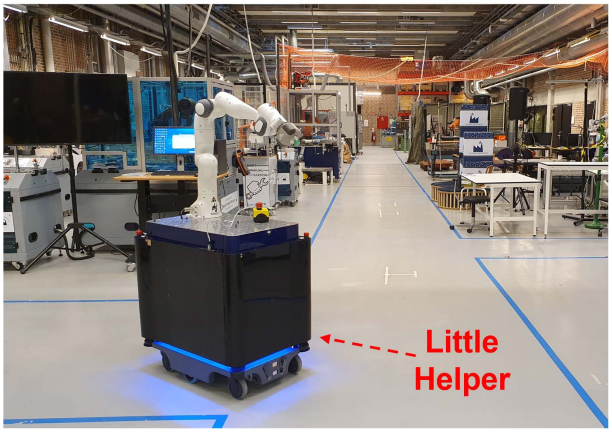


FIGURE 2. Our Little helper robot and the AAU learning factory.

generation [49]. LH combines a MiR 200 (on the bottom) and a Franka Emika Panda collaborative Robot (on the top), as seen in Fig. 2.

a: DIALOGUE TASK

An ontology of a dialogue task is formed by a domain, dialogue act type, and dialogue slots. Table 1 shows the defined ontology of the IRWoZ dataset.

As mentioned earlier, four domains are identified from the two scenarios (seen Appendix C).

- Delivery. A transportation task, where a mobile industrial robot delivers a package.
- Position. A positioning task, where a mobile industrial robot edits its position/location (e.g., add a new position) on a 2D shop floor map.
- Assembly. A product assembly task, where an industrial manipulator assembles a requested product.
- Relocation. A relocation task, where an industrial manipulator assists with the relocation of an object, e.g., grasping, moving.

There are three types of dialogue acts: database-related, task-related, and general greeting. The database search conditions are set based on extracted slots if a task requires database querying. For instance, a general delivery task requires a worker to specify the destination. The ToD system should be able to verify whether or not a particular destination (i.e., *location*) exists in the system database. There are two types of database act: required (*DB_request_req*) and optional (*DB_request_opt*) (e.g, name of the recipient for a delivery task) acts. The task-related dialogue act denotes additional information that a worker needs to offer to complete a task, e.g., *quantity* of an assembly task. It also includes required (*T_inform_req*) and optional(*T_inform_opt*) slots. The verification results of the required or optional dialogue acts are specified in the search results(i.e., search_results). Additionally, general greeting acts, like "*thank you*", "*goodbye*", are also provided.

Slots are required core information (e.g., product, area) extracted from a worker's utterance to accomplish a manufacturing task. There are a total of 19 slots identified throughout the above four domains. Domains may share the slots; for example, slot *color* can be used to specify the color of a product for an assembly task or the color of a package used in a delivery task.

TABLE 1. The ontology of IRWoZ dataset including the domain, dialogue act types, and slots.

Domain	Delivery, Position, Assembly, Relocation
Dialogue act	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 Search_results
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

- You are working on an internal transportation task.
- You work in the lab of the AAU Smart factory.
- Your colleague John is expecting a package from you.
- John is working at the warehouse in the lab.
- The package is a small yellow box.
- You use Little Helper to deliver the package.
- The area (lab) and location (warehouse) should be registered in the database.
- The following details of the package are optional:
  - recipient: John
  - size: small
  - Color: yellow

FIGURE 3. A sample task template for a delivery task.

b: ROLE OF PARTICIPANTS

Human-to-human method is leveraged to collect dialogue corpus regarding the above four domains. Each participant is introduced to the LH robot and instructed on how to operate LH to perform tasks. To collect data, one participant assumes the role of an industrial robot (i.e., the *wizard*), while the other one acts as a *shop floor worker*.

The *shop floor worker* is asked to randomly choose a task and initiate a dialogue. By following the same approach in [38], sampled task specification (e.g., Fig. 3) crossing four domain are distributed to *shop floor worker*. The *wizard* needs to respond to the *shop floor worker* according to the required task. The maximum conversational depth and the number of

TABLE 2. The "ARE" principles [50] and three conversation strategies [51] for generating humanised response with examples.

Principle	Definition	Examples	Strategy	Definition	Examples
Anchor	the conversation with a topic that is part of the 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?*

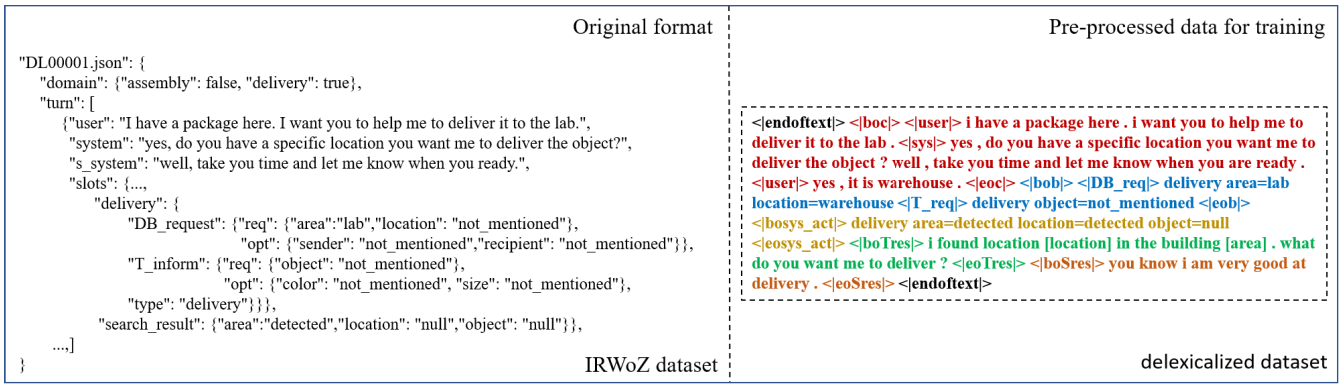


FIGURE 4. Sample corpora from raw IRWoZ dataset (left) and generated delexical data (right).

continuous dialogue turns for a task are not limited to creating a natural dialogue environment [51]. Since individuals structure their utterances differently even when assigned the same task, the *shop floor worker* is encouraged to compose the sentence uniquely.

The *wizard* is given access to the back-end database and a list of the robot-controlling APIs. Once the *shop floor worker* initiates a task-related conversation, the *wizard* responds appropriately. If the task requires database verification, the *wizard* needs to provide ground truth results based on the database. They then extract task-related slots from the *shop floor worker* 's utterance. To construct a humanized response for the ToD, the *wizards* are urged to order their utterances using their own words. If preferred, *wizards* are also supplied with work-related small chat principles and task-related human-to-human conversation strategies.

3) SMALL TALK RESPONSE GENERATION: ARE PRINCIPLES

Unlike [28], we do not employ pre-trained models (e.g., GPT-2 [52], BlenderBot [6]) to generate candidate chit-chat responses, annotate and filter them manually. Rather than that, we collect individual human responses to the IRWoZ dataset during the conversations. The observation of human collaboration demonstrates that most tasks-related conversations may also include small talk, which can be incorporated into the ToD process to build rapport, establish trust, and increase user engagement. [53], [54], [55]. Inspired by [50], Anchor, Reveal and Encourage (ARE) principles are introduced to assist the small talk response generation. Table 2 shows the descriptions of ARE and its associated examples. *Anchor* facilitates mutual understanding and builds a friendly dialogue that can lead to task development. For instance, a shop floor operator may ask the mobile robot to deliver a package to the warehouse where the main storage room is located. The ToD may respond by combining task-related and small talk responses (the italic part) by saying:" Sure, I will do that. *I know that place, it is a quite busy area.*" Establishing a trustworthy dialogue is likewise reliant on trust. Individuals trust those who share more personal information with them (i.e., *Revealing*). Automated assembly tasks may require the operator to instruct the robot to assemble an unfamiliar product. Therefore, instead of replying: "Sorry, I cannot.", ToD may say: "Sorry, I cannot. *it is quite new to me. I have not learned how to do that yet.*". Another way to enhance user engagement is to *encourage* them to speak and involve them in conversation. The ToD may give a response: "*Hey, how are you? Do you have a nice day?*" when the operator greets the robot: "Hey robot, good morning!".

4) TASK RELATED RESPONSE ENHANCEMENT: CONVERSATION STRATEGY

While the use of small conversation helps to increase user engagement, the primary objective of the developed ToD is to assist manufacturing tasks. Three human-to-human

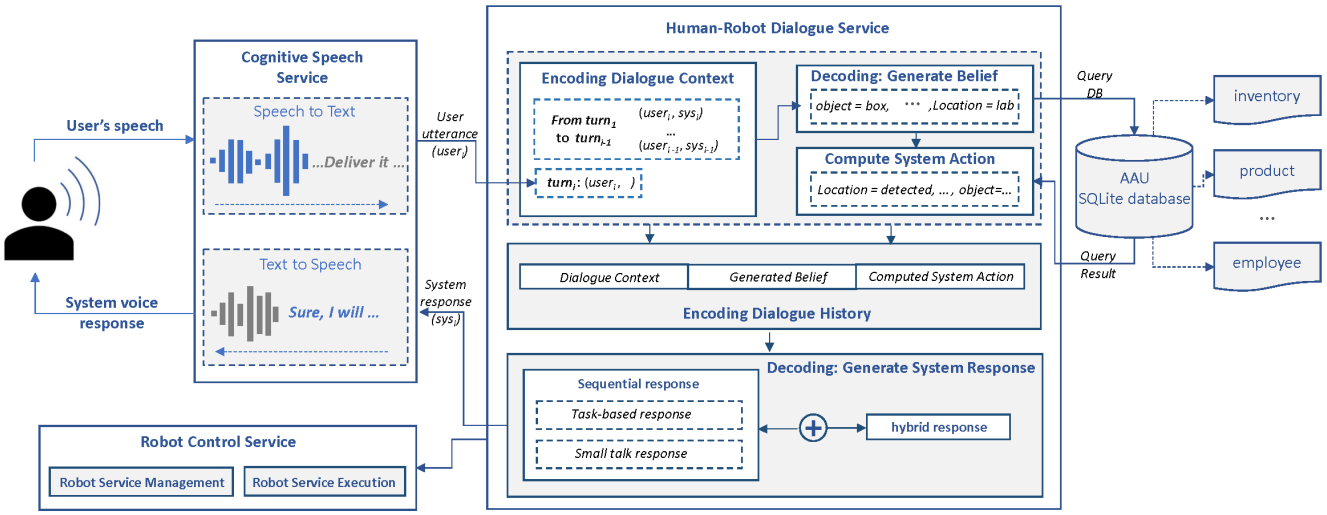


FIGURE 5. An overview of ToD4IR system architecture.

conversation strategies [51], *end topics with a suggestion*, *elicit more information* and *clarifying*, are introduced in order to improve task-completion rate while maintaining a more humanised and task-related response. Table 2 contains the definitions of the strategies and their accompanying examples.

The ToD is expected not to give a "yes" or "no" answer for a requested task but to the mining operator's intention to the greatest extent and provide alternative solutions to assist in completing the manufacturing task. For example, for a transportation task, the operator may accidentally give a location that is not registered in the system. The ToD may use *end topics with a suggestions strategy* and propose a *Position* task to the operator to mark the location first and say: "well, I don't know the place. Can you register it in the system first?". ToD should also be able to ask the operator to *elicit more information* if the instruction is not explicit. For example, the operator may forget to mention the required task-related information for an assembly task and ask: "Hey robot, can you assemble ten pieces of this?" The ToD may say: "Can you describe a bit more of the thing you want to assemble?" Clarification and confirmation are essential for task execution. Operators may alter tasks during the conversation, e.g., aborting a task-switching a task. For example, if the operator decides to abort a task during its execution, ToD needs to understand if that is what the operator wants or just a misunderstanding. Therefore, ToD should be able to use the *Clarifying* strategy to ask: "Sorry, do you mean you want to abort the task?".

5) DATA STRUCTURE

To maintain high scalability, IRWoZ uses a data structure similar to that of the most popular MultiWOZ 2.0 [38]. Each dialogue is divided into two sections: the domain and the turn. The turn section includes multiple dialogue turns, including user and system utterances, belief states, and system actions. The dialogue is saved in JSON format. The left side of Fig. 4 illustrates an example of IRWoZ data. To extend the IRWoZ, the user needs to follow the current data structure and add new JSON elements which includes the desired domain, dialogue turns, database-related slots, task-related slots, and search results.

B. SYSTEM ARCHITECTURE

The proposed ToD4IR comprises three services: cognitive speech, human-robot dialogue, and robot control. Fig. 5 illustrates a high-level system architecture of ToD4IR.

1) COGNITIVE SPEECH SERVICE

The ToD4IR is designed in such a way to accept human speech as input and provide near-human voice as the output to create a natural and flexible communication environment.

Generally, most popular conversational AIs support a trigger word, e.g., *Ok Google*, *Alexa*, for activating the voice service. The service will terminate if no active human voice is detected within a certain period. Though most smart speakers adopt this interaction strategy, such interaction is not as natural as human-to-human communication. The experience of real-world human-to-human interaction demonstrates that there are no universal standards for constructing dialogues. The work-related dialogue among participants might be continuous or discrete, with a particular greeting at the beginning of the conversation. Furthermore, observation from our previous work [10] also shows that inaccurate human intent prediction is another concern, as a portion of a human's speech may fall between two continuous voice detection periods. Therefore, the voice interface of ToD4IR is designed to keep listening to the human voice in the background rather than manually invoking the service using trigger words. Thus, ToD4IR can provide a near-human ability to participate in a conversation at any moment without requiring the user

to employ trigger words, as long as the task-related user utterance is identified.

Two cognitive speech services of Microsoft,[4] speech-to-text (STT) and text-to-speech(TTS), are leveraged to convert streamed human voice to the transcript and generate the near-human voice as a response, respectively.

2) HUMAN-ROBOT DIALOGUE SERVICE

As the core service of ToD4IR, the human-robot dialogue service (HRDS) is composed of five components: encoding dialogue context, decoding belief state, computing system action, encoding dialogue history, and decoding system response. SQLite database,[5] which stores real-time production data, is leveraged to assist grounding response generation.

3) TASKS OF THE HRDS

There are four tasks, data preprocessing, belief state prediction, system actions generation, and system response generation, running on the HRDS.

a: TASK 0: IRWOZ DATASET PREPROCESSING

We define the dialogue context as follows:

$$C = \{U_1, Sys_1, \ldots, U_i, Sys_i\} \tag{1}$$

where U_i represents the user's utterance at turn i, and Sys_i is defined as

$$Sys_i = \begin{cases} Tres_i \oplus Sres_i, & if\ Sres_i \neq \emptyset \\ Tres_i, & otherwise \end{cases} \tag{2}$$

where $\oplus$ denotes text concatenation, *Tres_i* and *Sres_i* represent task-related and small talk-related responses, respectively. As aforementioned, the *wizard* may either use a hybrid response (i.e., only *Tres_i*, self-organized natural response which might mix small talk response with the task response) or concatenates *Tres_i* and *Sres_i* (guided by ARE principles and human-to-human conversation strategies). Dialogue belief B defines the dialogue states:

$$B = \{b_1, b_2, \ldots, b_i\} \tag{3}$$

where b_i denotes dialogue state at turn i, and b_i defined as

$$b_i = DB\_req_i \oplus DB\_opt_i \oplus T\_req_i \oplus T\_opt_i \tag{4}$$

where DB_req_i and DB_opt_i indicate the required and optional belief states for database retrieval at turn i, T_req_i and T_opt_i for the required and optional task related belief states. Both DB_opt_i and T_opt_i can be NULL if operator does not provide them. System actions, Sys_act, define verified results from the operator's utterance:

$$Sys\_act_i = DB\_SR_i \oplus T\_SR_i \tag{5}$$

where DB_SR_i and T_SR_i denote the generated system actions at turn i based on database search results and state tracking results, respectively. Additionally, the system response, Res, is defined as:

$$Res_i = Sys_{i+1} \tag{6}$$

where the response, Res_i, at turn i will be the part of the dialogue context, Sys_{i+1}, at turn $i+1$.

We define a dialogue, D, of HRDS as:

$$D = \{C, B, Sys\_act, Res\} \tag{7}$$

Therefore, the training dataset, as raw IRWoZ (see the original format on the left side of Fig. 4), needs to be reconstructed and annotated by following the above dialogue structure before moving to HRDS. Table 3 defines the special tokens used to identify components of the raw text. One of the tasks of data preprocessing is to extract and annotate the raw text from IRWoZ and generate the expected dialogue dataset (see the right side of Fig. 4).

TABLE 3. Special tokens used to identify components of the text.

Token	description
<|endoftext|>	indicate beginning and end of text
<|boc|>	indicate beginning of the context
<|eoc|>	indicate end of the context
<|bobl|>	indicate beginning of the dialogue belief state
<|eobl|>	indicate end of the dialogue belief state
<|user|>	indicate user's utterance
<|sys|>	indicate final system response
<|DB_req|>	indicate required slots for database query
<|DB_opt|>	indicate optional slots for database query
<|T_req|>	indicate required slots for target task
<|T_opt|>	indicate optional slots for target task
<|bosys_act|>	indicate beginning of the system actions
<|eosys_act|>	indicate end of the system actions
<|boTres|>	indicate beginning of the task-oriented response
<|eoTres|>	indicate end of the task-oriented response
<|boSres|>	indicate beginning of the small talk response
<|eoSres|>	indicate end of the small talk response

b: TASK 1: PREDICT BELIEF STATE

ToD4IR follows a similar strategy, such as [9], [23], which leverages the autoregressive model (see the III-B4) to generate the belief states, task-related response, and small talk-related response stepwise. The joint probability, $p(x)$, over the given sequence of the text of a dialogue can be factorized as:

$$p(C_i) \rightarrow p(B_i|C_i) \rightarrow p(Res_i|(C_i, B_i, Sys\_act_i)) \tag{8}$$

Dialogue belief, b_i, at turn i is predicted based on the dialogue context, $\{U_1, Sys_1, \ldots, U_{i-1}, Sys_{i-1}, U_i\}$, from turn 1 to turn i (only user utterance at turn i). The predicted belief states that b_i, is a sequence of text which is formed as follows:

$$b_i = \{T\ b\_s_1 = b\_v_1, \ldots, b\_s_m = b\_v_m\}, \tag{9}$$

where T represents the predicted domain, b_s and b_v stand for the belief slot and the slot value, respectively, and m is the total number of the predicted belief slots. b_s should belong to the set of predefined slots, slt_T, of the desired task T,

$$slt_T = \{slt_1, slt_2, \ldots, slt_n\} \tag{10}$$

[4]https://azure.microsoft.com/en-us/services/cognitive-services/
[5]https://www.sqlite.org/index.html

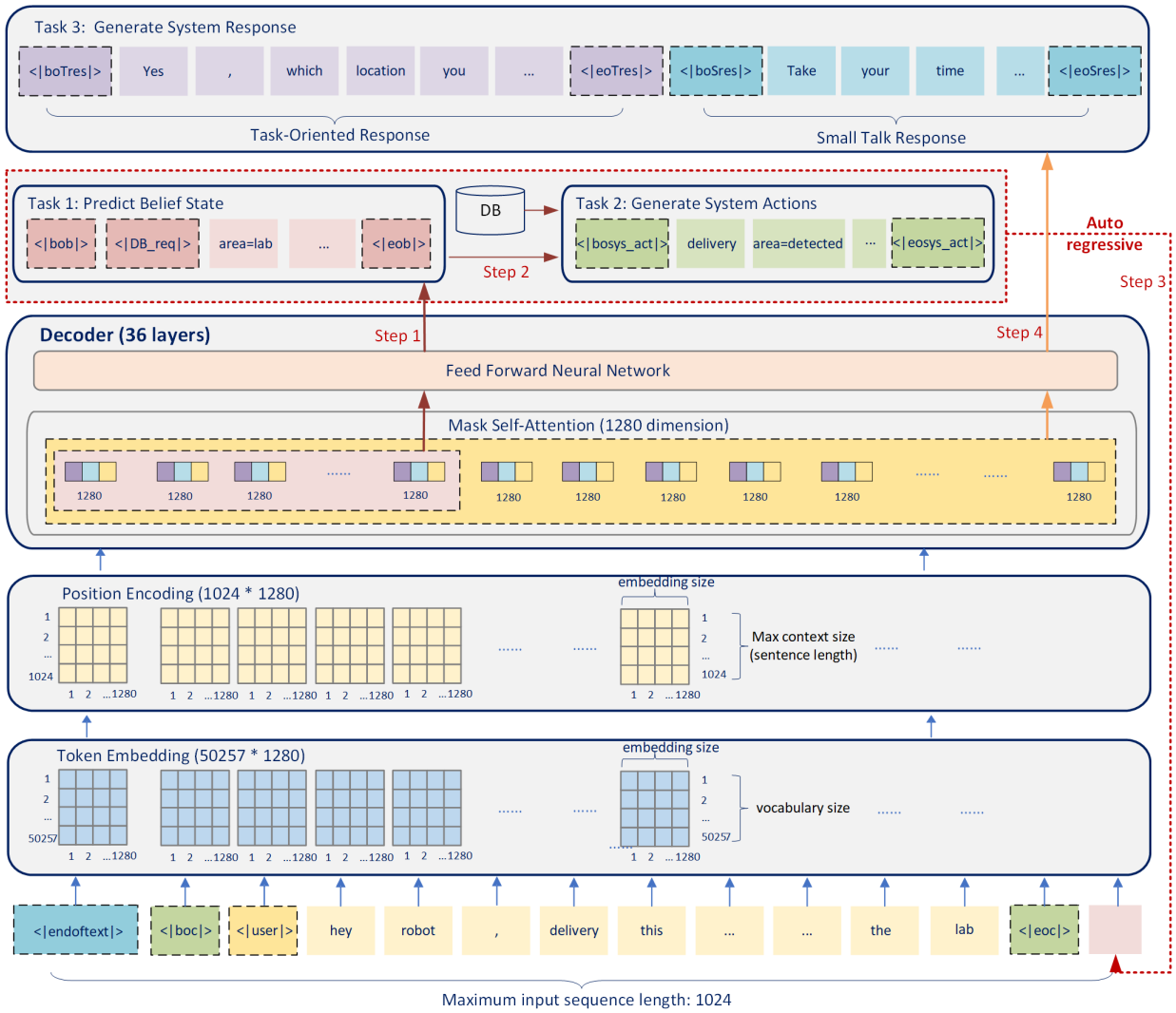


FIGURE 6. **An overview of the ToD4IR model architecture including three tasks, *predict belief state*, *generate system actions* and *generate system response*.**

$$b\_s_i \in slt_T, 1 \leq i \leq m \tag{11}$$

where n is the total number of the task slots S and slt_i is a slot related to the database or the task. The training goal of belief prediction, $\mathcal{L}(B_i^T)$, is defined as:

$$\sum_{j=1}^{o} logp\theta(b\_s_j|(b\_s_1, \ldots, b\_s_{j-1}), C_i^T)), \tag{12}$$

where o is the total number of slots of a belief state sequence, and θ is the learning neural network parameters. If the slt_j is an optional slot (i.e., DB_opt or T_opt), it will not be predicted unless it is detected from operator's utterance.

c: TASK 2: GENERATE SYSTEM ACTIONS

System actions, Sys_act, include database-related actions and task-related actions. If database querying is required based on the predicted belief states, the corresponding database-related slots of belief states are extracted as query parameters. The extracted task-related slots and the database querying results are mapped to system actions (i.e., T_act and DB_act) and are defined as follows:

$$DB\_act, T\_act \in \{null, detected, undetected\} \tag{13}$$

where *null* means the operator does not provide the slot while it can be *detected* from the operator's utterance (if it is a task-related slot) or matched with database query results (if it is a database-related slot). Naturally, *undetected* means that the slot remains undetected.

d: TASK 3: GENERATE SYSTEM RESPONSE

The delexicalized task related system response (e.g., text between '|boTres|' and '|eoTres|' of the right side of Fig. 4), and small talk response (e.g., text between '|boSres|' and '|eoSres|' of the right side of Fig. 4). The goal of training,

$\mathcal{L}(Tres_i^T)$ and $\mathcal{L}(Sres_i^T)$, are as follows:

$$\sum_{i=1}^{N_k} logp\theta(Tres_k|Tres_{<k}, C_i^T, B_i^T, Sys\_act_i^T), \quad (14)$$

$$\sum_{i=1}^{N_j} logp\theta(Sres_j|Sres_{<j}, C_i^T, B_i^T, Sys\_act_i^T, Tres_i), \quad (15)$$

where N_k and N_j represent the length of the sequence of task-related and small talk-related responses, respectively. $\mathcal{L}(Sres_i^T)$ is not calculated if operator mix the small talk response with task-related response (see Fig. 5, hybrid response).

4) BACKBONE MODEL

ToD4IR, like previous state-of-the-art (SoTA) works (e.g., SOLOIST [23], SimpleToD [9] and MinTL [22], is implemented using two auto-regressive pre-trained language models, GPT-2 and GPT-Neo [56]. Those models were trained on massive volumes of open Web material and learned how to complete a sentence in a given context. Such models are largely used in downstream NLP tasks (e.g., machine translation, answering questions, text generation in our case), where they are fed a small task-specific dataset for fine-tuning the desired final model. We briefly introduce GPT-2 and GPT-Neo, which are used in this paper.

a: GPT-2

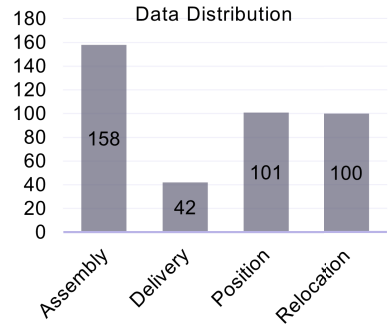
The OpenAI[6] GPT-2 follows decoder only transformer [57] architecture in a self-supervised manner. It is trained with causal language modeling (CLM) with the target of generating coherent text based on the previously given text. Hugging face[7] provides five different sizes, distilgpt-2, gpt-2, gpt2-medium, gpt2-large and gpt2-xl. Different from [8], [9], [23], ToD4IR uses gpt2-large (with 36 layers and 1280M parameters) instead of gpt-2 as one of the backbone models in our case (see Fig. 6).

b: GPT-NEO

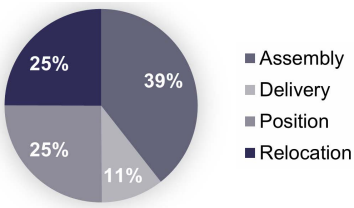
Since the latest pre-trained language model, GPT-3 (with 175B parameters), has not yet been open-sourced, EleutherAI's GPT-Neo is leveraged as our second backbone model. GPT-Neo is an open-source transformer model which replicates the GPT-3 architecture. It is trained on the Pile [58] dataset, which has an 825GB English text corpus. Evaluation results of linguistic reasoning and physical and scientific reasoning show the GPT-Neo (with 2.7B parameters) resembles GPT-3 in performance.[8]

5) ROBOT CONTROL SERVICE

Based on our previous work [10], ToD4IR is designed to be agnostic of robot hardware. The robot control service is implemented as a pair of components: a robot service management (RSM) and a robot service execution (RSE). The output of the HRDS (i.e., task-related information and commands) will be sent to RSE to ground language instructions into robot actions, which invokes robot control functionalities (e.g., package delivery) via communication protocols (e.g., TCP/IP). RSM periodically synchronizes, once per week by default, the local robot services and skills with the ToD4IR server to update the robot services/skills registered on the local client-side and the robot controlling scripts.



(a) Distribution of dialog corpora span four domains.



(b) % of the total dataset

FIGURE 7. Composition and percentage of dialogue dataset.

IV. EXPERIMENTS AND DISCUSSIONS

This section evaluates the proposed approach to address two research questions: Q1: Given that ToD4IR is a conversational AI for industrial robots, how does ToD4IR perform tasks within the four defined domains? Q2: Is the ToD4IR able to augment user experience through embedded small talk?

A. EXPERIMENTAL SETUP

We trained and evaluated our ToD4IR on Aalborg University Cloud, where each server is configured with Intel(R) Xeon(R) Gold 5118 (12 cores), 256GB of memory, and Nvidia V100 with 32GB VRAM. The ToD4IR is implemented based on HuggingFace Transformers 4.7.0,[9] Microsoft DeepSpeed 0.4.0[10] and Torch 1.7.[11]

[6]https://openai.com/blog/better-language-models/
[7]https://huggingface.co/gpt2
[8]https://github.com/EleutherAI/gpt-neo
[9]https://huggingface.co/docs/transformers/index
[10]https://github.com/microsoft/DeepSpeed
[11]https://pytorch.org/get-started/locally/

TABLE 4. The components of the annotated dialogue corpus of IRWoZ dataset. ToD4IR trains on a sequence of such annotated dialogue corpus.

Component	Description	Example
Dialogue Context	It includes user utterance and system response.	<|boc|> <|user|> could you deliver the box for me please ? <|sys|> sure , where do you want to deliver it? <|user|> yes, it is the office in the lab. <|eoc|>
Belief State	It contains database-related slots, task-related slots (including database and task-related), and the user's dialogue act.	<|bob|> <|DB_req|> delivery area=lab location=office <|T_req|> delivery object=box <|eob|>
System Actions	It comprises the database search results and task-related slots extraction.	<|bosys_act|> delivery area=detected location=detected object=detected <|eosys_act|>
Task response	It responds to the user's spoken command.	<|boTres|> ok , i found location [location] in the [area] . <|eoTres|>
Small Talk response	It is a casual style of discourse that omits any necessary functional themes of a task.	<|boSres|> i am so happy!<|eoSres|>

1) PREPARATION OF DATASETS

While the WoZ method enables a more flexible and straightforward generation of low-noise dialogue corpora, the time-consuming data collection and annotation processes need high accuracy. The primary difficulty lies in checking whether the appropriate slots are given, validating whether the slots match database search results, and manually annotating slots. Fig. 4 shows an example of raw IRWoZ dialogue and the respective annotations.

To address this issue, a Flask-based[12] IRWoZ web application is constructed. The application features two interfaces, one for the user and one for the Wizard. The user submits their utterance using the online form, and the Wizard automatically detects and checks the extracted dialogue acts (e.g., delivery, position) via the web form. The Wizard is given two distinct input text fields to deliver task-related and small-talk-related responses. When the user confirms the chat is complete, the web application automatically annotates and saves the conversations. We provide an example of the dialogue collection and annotation processes using the web application developed in the appendix B.

Eighteen individuals have been asked to participate in the gathering procedure for the IRWoZ dataset. The dataset contains 158/42/101/100 dialogues corpus span over assembly/delivery/position/relocation (see Fig. 7), with each dialogue having at least two turns and augmenting over 88.7% of system responses with small talk. Table 4 summarizes the essential components of the annotated dialogue with examples. To evaluate the proposed ToD4IR on our IRWoZ dataset, we divide it into 60%, 20%, and 20% for training, validation, and testing, respectively.

2) AUTOMATIC EVALUATION

To respond to the first research question, we train our ToD4IR based on GPT architecture with five pre-trained language models: gpt2, gpt2-large, gpt2-xl, GPT-Neo (1.3B), and GPT-Neo(2.7B). The ToD4IR follows the end-to-end dialogue pattern in which the evaluation mainly involves two aspects, dialogue state tracking, and system actions and response generation [38].

[12]https://flask.palletsprojects.com/en/2.0.x/

TABLE 5. Evaluation of dialogue state tracking on IRWoZ using joint accuracy and slot accuracy metrics.

Model	Joint Goal Accuracy	Slot Accuracy
ToD4IR-gpt2	0.789	0.941
ToD4IR-gpt2-large	0.856	0.959
ToD4IR-gpt2-xl	0.861	**0.964**
ToD4IR-gpt-neo (1.3B)	0.845	0.957
ToD4IR-gpt-neo (2.7B)	**0.866**	0.963

In order to evaluate the performance of ToD4IR, we used three metrics.

- 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.
- Bilingual evaluation understudy (BLEU) [59]. It is mainly used for measuring the fluency of the generated text.

Among the above three evaluation metrics, joint goal accuracy and slot accuracy are commonly used in dialogue state tracking tasks, and BLEU is leveraged for response generation.

a: DIALOGUE STATE TRACKING

This task aims to assess ToD4IR's ability to predict dialogue state in the given dialogue context, which includes domain, slot, and value. Table 5 compares the ToD4IR's joint objective and slot accuracy to that of various backbone models on the IRWoZ dataset. ToD4IR-gpt2-xl outperforms other models at 0.964 of slot accuracy, whereas ToD4IR-gpt2-neo(2.7B) outperforms at 0.866 of joint goal accuracy.

b: SYSTEM ACTIONS AND RESPONSE GENERATION

In this task, ToD4IR should generate system actions and responses given the ground truth dialogue states and database search results. Due to the WoZ data collection method, the ground truth of database search results is manually incorporated into the system actions of the IRWoZ dataset.

TABLE 6. Context-to-response evaluation on IRWoZ.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
ToD4IR-gpt2	0.4812	0.3882	0.3386	0.2968
ToD4IR-gpt2-large	**0.6013**	**0.5349**	**0.5032**	**0.4763**
ToD4IR-gpt2-xl	0.5709	0.5059	0.4727	0.4432
ToD4IR-gpt-neo (1.3B)	0.5183	0.4410	0.4001	0.3645
ToD4IR-gpt-neo (2.7B)	0.5599	0.4853	0.4469	0.4140

As a result, ToD4IR eliminates the requirement for database searches during the training and validation processes. Table 6 compares the ToD4IR's BLEU 1-4 scores on the IRWoZ dataset's five models. At 0.6013, 0.5349, 0.5032, and 0.4763, ToD4IR-gpt2-large outperforms other models.

3) HUMAN EVALUATIONS

Human assessments provide a complete picture of response-generating performance, particularly in the generation of human-like response [60]. We leverage the human evaluation questions from [8], which cover engaging, interesting, human-like, and knowledgeable, for response generation assessments.

As mentioned, 20% dialogue corpus is selected as the test data set. We feed those data samples to the five models, ToD4IR-gpt2, ToD4IR-gpt2-large, ToD4IR-gpt2-xl, ToD4IR-gpt-neo (1.3B) and ToD4IR-gpt-neo(2.7B) to obtain the responses. Task-related responses and small talk responses are highlighted in each of the dialogues.

To verify the validity of the assessment findings, domain specialists within robotics, computer science, linguistics, and humanities and various end-users such as shop floor workers and lab engineers were invited to serve as evaluators. The evaluators are asked to rate their answers on a predetermined scale.[13,14] Each number on the scale denotes a different quality level, ranging from *Not at all* to *Absolutely*. For instance, the inquiry *How much would you prefer to talk to the ToD4IR?* is used to determine whether the user is engaged and willing to speak with the ToD4IR.

In some circumstances, particularly when evaluating the engaging and knowing competence, the dialogue context is essential to evaluate the generated system response. However, the reference context is not required to evaluate fluency (without grammatical errors), interest, and human likeness. We give context for each generated response in our evaluation (Appendix C presents six examples and the web interface for online evaluation).

We showcase each model's individual score, in Table 7, and summarize how engaging, interesting, human-like, and knowledgeable the dialogue was. Table 8 reports the overall score of ToD4IR.

B. DISCUSSION

1) DATASET

Although the IRWoZ is being offered as the first industrial-oriented conversation dataset for human-robot interaction in the manufacturing sector, it is a small-scale dialogue corpus with limited domain coverage. Additionally, an imbalanced data distribution (i.e., delivery occupies only 11% of the conversation corpus) affects ToD4IR's performance on the delivery job.

[13]https://shorturl.at/cmES1

[14]We obtained the ethics approval from Aalborg University regarding the online questionnaire.

TABLE 7. Human evaluation results of the generated response of ToD4IR.

Model	Engaging	Interesting	Humanlike	Knowledgeable
ToD4IR-gpt2	**50%**	25%	**37.5%**	37.5%
ToD4IR-gpt2-large	12.5%	-	12.5%	-
ToD4IR-gpt2-xl	12.5%	**50%**	-	12.5%
ToD4IR-gpt-neo (1.3B)	-	-	12.5%	-
ToD4IR-gpt-neo (2.7B)	25%	25%	37.5%	**50%**

TABLE 8. Human evaluation results of overall engaging, interesting, human-like, and knowledgeable score.

	Engaging	Interesting	Humanlike	Knowledgeable
Score	8.0	8.5	7.0	8

2) NOISE-LABELS

In comparison to other datasets, such as MultiWoZ 2.0, IRWoZ contains far fewer mis-annotations, which adds to the high accuracy of dialogue state tracking (see table 5). The first reason is that the IRWoZ is compiled by 18 individuals, including students, academics, and shop floor workers with backgrounds ranging from robotics to computer science. As a result, the corpora of dialogue collected are significantly more pure and professorial. Second, a web application (source code can be found on our Github) is created to support the data collection. The user may directly engage with the application with the four domain tasks. The application's back end automatically annotated and verified each dialogue, including annotations and database search results. Additionally, human verification is performed as the last step, ensuring that the collected dialogue corpora contain fewer mis-annotations than other datasets.

3) DIALOGUE RESPONSE

One of the objectives of our method is to train the ToD4IR to generate more natural and human-like responses. As a result, natural human responses are expected during the dialogue simulation. However, examining the collected data, we realize that when users communicate with the system, they use a less complicated version of their language. For instance, a simple response of "*Ok*." is typically used when they affirm a system-generated inquiry, whereas "*Yes, I believe so*." or "*yes, you are correct*." is frequently used when they speak with a human. The participants confirm this trend during the collection of data. The other observation is that ten out of the eighteen persons prefer to structure their small talk responses instead of directly using the ARE principles.

4) PRE-TRAINED MODELS

The ToD4IR is mainly driven by gpt2-large and gpt-neo(2.7B) models with pre-trained weights. Each model is

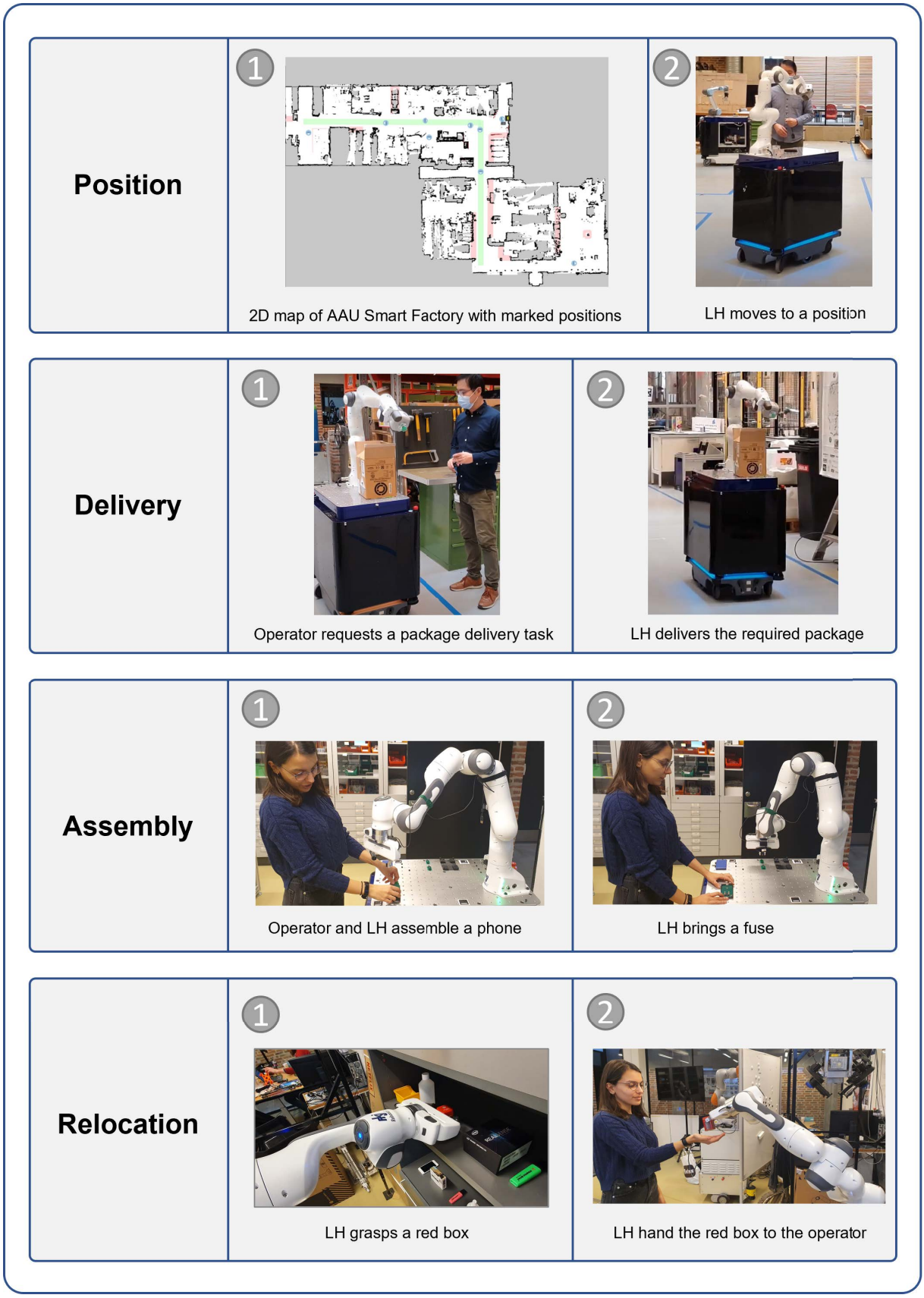
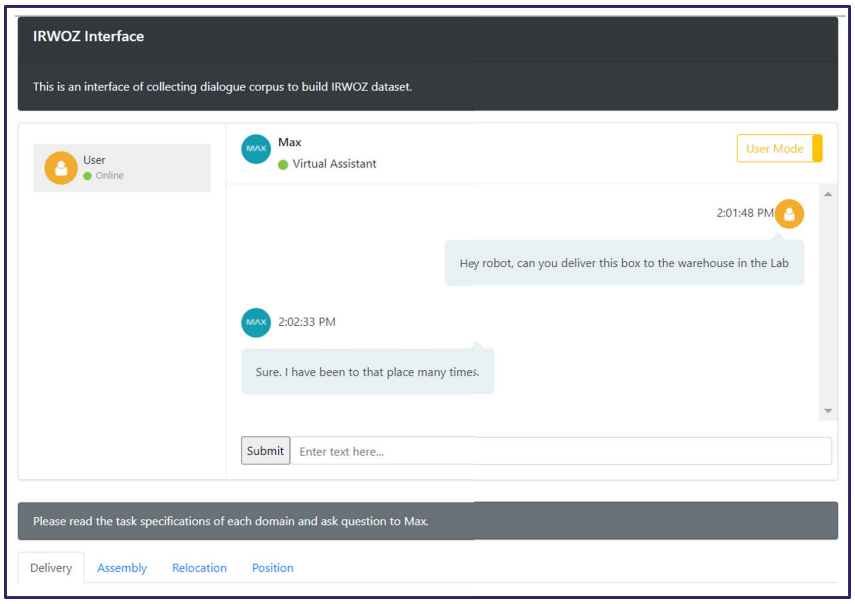
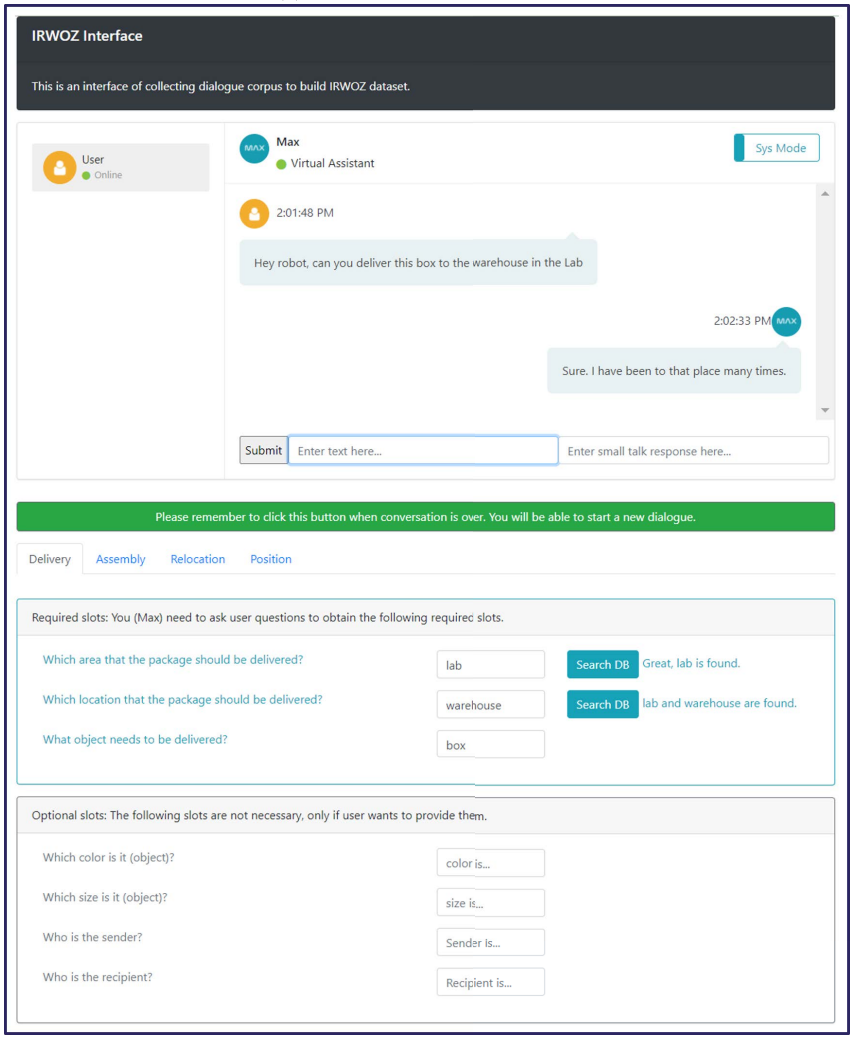


FIGURE 8. Identified four domains for ToD4IR.



(a) Interface from the user side



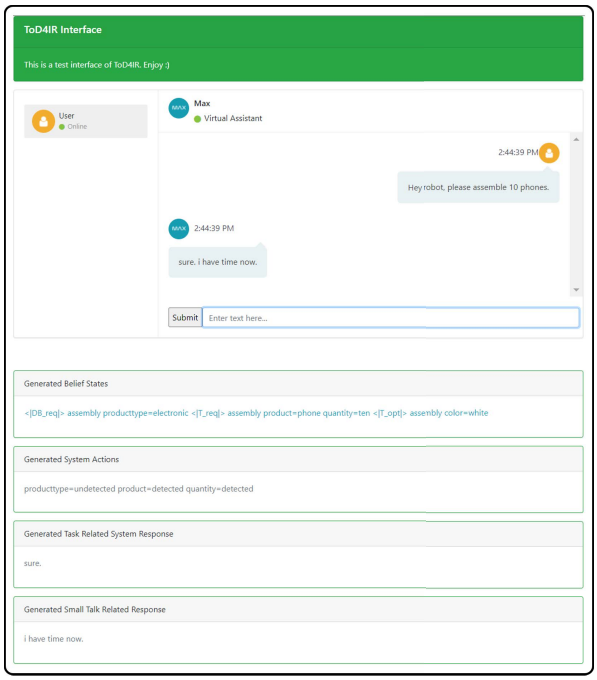
(b) Interface from the Wizard side

FIGURE 9. **Web application for IRWoZ dialogue corpus collection.**

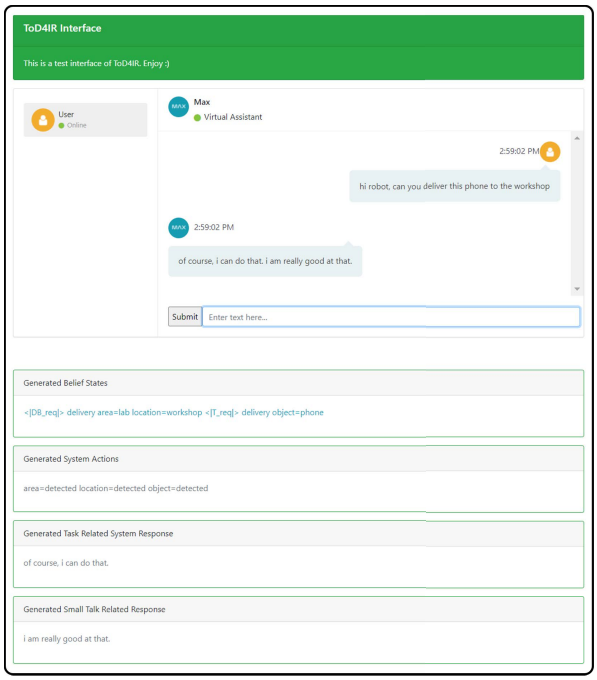
TABLE 9. ToD4IR end-to-end response generation on the test set of IRWoZ. We show the generated task-related and small talk-related responses by different models.

Example 1	**Context**	**<|boc|> <|user|> Hi guy! Can you assemble 270 legos for me please? <|eoc|>**
Model	**Task Response**	**Small Talk Response**
ToD4IR-gpt2	<|boTres|> Sure. What kind of them? <|eoTres|>	<|boSres|> By the way, you are waiting in a long line. Please be patient. <|eoSres|>
ToD4IR-gpt2-large	<|boTres|> No problem. What kind of lego? <|eoTres|>	<|boSres|> I like lego very much. <|eoSres|>
ToD4IR-gpt2-xl	<|boTres|> No problem. What kind of [product]? <|eoTres|>	<|boSres|> I like lego very much. <|eoSres|>
ToD4IR-gpt-neo(1.3B)	<|boTres|> No problem. What kind of [product]? <|eoTres|>	<|boSres|> I like lego very much. <|eoSres|>
ToD4IR-gpt-neo(2.7B)	<|boTres|> No problem. What kind of lego? <|eoTres|>	<|boSres|> I like lego very much. <|eoSres|>
Example 2	**Context**	**<|boc|> <|user|> I need you to run down the workbench. <|eoc|>**
Model	**Task Response**	**Small Talk Response**
ToD4IR-gpt2	<|boTres|> Sure. <|eoTres|>	<|boSres|> I have been to that place many times. <|eoSres|>
ToD4IR-gpt2-large	<|boTres|> Yes. I knew that place. <|eoTres|>	<|boSres|> <|eoSres|>
ToD4IR-gpt2-xl	<|boTres|> Yes, I have been to that place many times. <|eoTres|>	<|boSres|> Is there anything else that I can do for you? <|eoSres|>
ToD4IR-gpt-neo(1.3B)	<|boTres|> Sure. <|eoTres|>	<|boSres|> I am glad to help <|eoSres|>
ToD4IR-gpt-neo(2.7B)	<|boTres|> Yes. <|eoTres|>	<|boSres|> I have been to that place many times. <|eoSres|>
Example 3	**Context**	**<|boc|> <|user|> Let's go to the maker space, robot <|sys|> well, where is that? it is quite a new place to me <|user|> OK, you can go to the basement instead <|eoc|>**
Model	**Task Response**	**Small Talk Response**
ToD4IR-gpt2	<|boTres|> Sure <|eoTres|>	<|boSres|> I am glad to help <|eoSres|>
ToD4IR-gpt2-large	<|boTres|> Sure, let me check it out. <|eoTres|>	<|boSres|> Great! <|eoSres|>
ToD4IR-gpt2-xl	<|boTres|> No problem <|eoTres|>	<|boSres|> I am glad to help <|eoSres|>
ToD4IR-gpt-neo(1.3B)	<|boTres|> No problem <|eoTres|>	<|boSres|> I am in a familiar place there. <|eoSres|>
ToD4IR-gpt-neo(2.7B)	<|boTres|> Sure. <|eoTres|>	<|boSres|> Let me know if you need other help <|eoSres|>
Example 4	**Context**	**<|boc|> <|user|> Can you reach the injection molding machine <|eoc|> <|bob|>**
Model	**Task Response**	**Small Talk Response**
ToD4IR-gpt2	<|boTres|> Em, I can not. <|eoTres|>	<|boSres|> It is a difficult task for me <|eoSres|>
ToD4IR-gpt2-large	<|boTres|> No. This is beyond my ability <|eoTres|>	<|boSres|> Let me know if you need any other help. <|eoSres|>
ToD4IR-gpt2-xl	<|boTres|> No, it is not possible <|eoTres|>	<|boSres|> but I can bring other things instead. <|eoSres|>
ToD4IR-gpt-neo(1.3B)	<|boTres|> No. <|eoTres|>	<|boSres|> I do not know that. <|eoSres|>
ToD4IR-gpt-neo(2.7B)	<|boTres|> No. This is quiet new to me <|eoTres|>	<|boSres|> but I can sure do that. <|eoSres|>
Example 5	**Context**	**<|boc|> <|user|> Hi friend! can you assemble eight blue phones? <|sys|> Of course. When do you want them? Look, I still have one parcel at hand! Maybe you need to wait. <|user|> All right. I can wait. So can you assemble them in the afternoon? <|eoc|>**
Model	**Task Response**	**Small Talk Response**
ToD4IR-gpt2	<|boTres|> All right. I will finish them in the afternoon. <|eoTres|>	<|boSres|> <|eoSres|>
ToD4IR-gpt2-large	<|boTres|> Sure. I must finish them before that. <|eoTres|>	<|boSres|> Trust me! <|eoSres|>
ToD4IR-gpt2-xl	<|boTres|> Sure. I must finish them before that. <|eoTres|>	<|boSres|> Trust me! <|eoSres|>
ToD4IR-gpt-neo(1.3B)	<|boTres|> All right. <|eoTres|>	<|boSres|> I am free now. <|eoSres|>
ToD4IR-gpt-neo(2.7B)	<|boTres|> Sure. I must finish them before that. <|eoTres|>	<|boSres|> but you know I have a good sense of time. <|eoSres|>
Example 6	**Context**	**<|boc|> <|user|> OK, have you been to the lego smart production line? <|eoc|>**
Model	**Task Response**	**Small Talk Response**
ToD4IR-gpt2	<|boTres|> Well, I do not know that. Can you register it in the system first? <|eoTres|>	<|boSres|> <|eoSres|>
ToD4IR-gpt2-large	<|boTres|> No, I have not been to that place. <|eoTres|>	<|boSres|> Anything else i can help you with? <|eoSres|>
ToD4IR-gpt2-xl	<|boTres|> No, I do not know that place. <|eoTres|>	<|boSres|> Anything else I can help you with? <|eoSres|>
ToD4IR-gpt-neo(1.3B)	<|boTres|> No, can you register it in the system first? <|eoTres|>	<|boSres|> I have not been to that place <|eoSres|>
ToD4IR-gpt-neo(2.7B)	<|boTres|> No, but I have not been to that place. <|eoTres|>	<|boSres|> Anything else I can do for you? <|eoSres|>

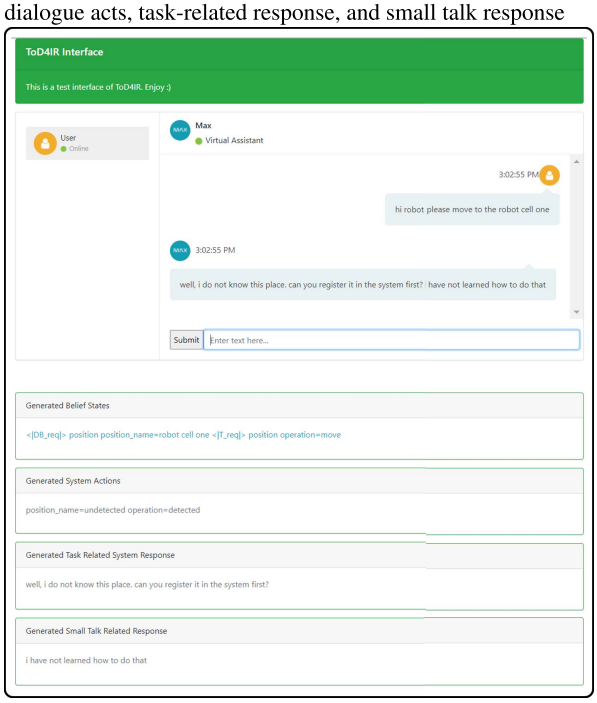
trained for 20 epochs. However, the largest model, ToD4IR-gpt-neo(2.7B), only surpasses ToD4IR-gpt2-large by a factor of 0.01 on joint goal accuracy while around 0.05 less on BLEU 1-4. Additionally, we compare the ToD4IR against three other gpt type models. The overall evaluation results indicate that shallow neural networks outperform deep



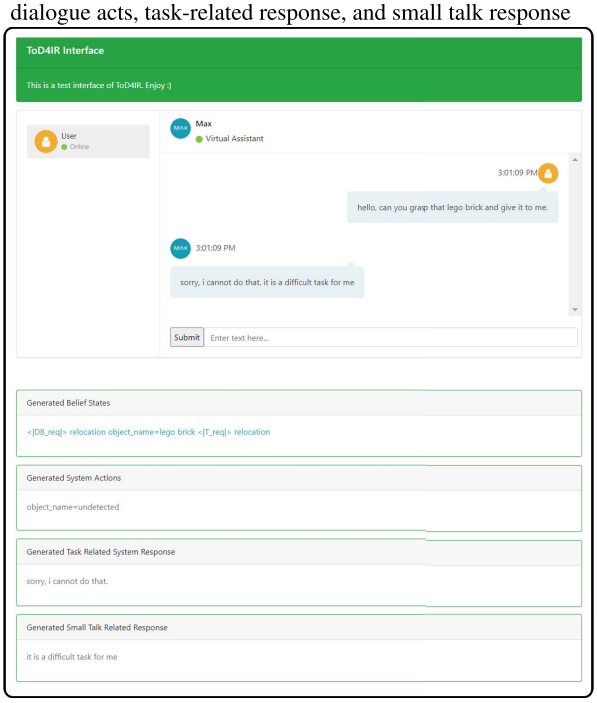
(a) An example of assembly task: generated dialogue state, dialogue acts, task-related response, and small talk response



(b) An example of a delivery task: generated dialogue state, dialogue acts, task-related response, and small talk response



(c) An example of position task: generated dialogue state, dialogue acts, task-related response, and small talk response



(d) An example of relocation task: generated dialogue state, dialogue acts, task-related response, and small talk response

FIGURE 10. **Text-based web application for testing ToD4IR.**

neural networks when ToD4IR is trained on a small-scale dataset.

V. CONCLUSION

In this study, we present ToD4IR, a humanized task-oriented dialogue system aimed at industrial robots. The first industrial-oriented dialogue corpus, IRWoZ, is constructed with 401 dialogues spanning four industrial tasks: delivery, assembly, position, and relocation. To aid ToD4IR in generating natural and humanized responses, we use small talk principles, known as ARE, along with human-to-human conversation strategies. Driven by a task-based neural network GPT, ToD4IR can predict the dialogue state and generate system actions using real-time database search results.

Additionally, it can generate work completion and user experience enhancement responses that include task-related (based on user goals) and small talk-related responses. Experiments demonstrate that ToD4IR achieves high accuracy in the dialogue state tracking and fluency in the generated response.

We hope that the proposed IRWoZ will inspire the dialog research community and industrial partners to continue investigating language-assisted human-robot interaction in manufacturing, contributing dialog corpora for new industrial domains, and fine-tune pre-trained pre-trained ToD4IR for new industrial tasks.

We intend to conduct extensive user research in the future to gather input on the naturalness and coherence of ToD4IR-generated dialogue and responses when COVID-19 constraints are removed.

APPENDIX A
IDENTIFIED FOUR DOMAINS FOR ToD4IR

Fig. 8 shows the identified four domains, position, delivery, assembly, and relocation, and the corresponding scenario examples. The HRI dialogue corpus is collected based on the tasks from the four domains.

APPENDIX B
WEB APPLICATION - IRWoZ DATA COLLECTION

Fig. 9 displays the designed web interface for collecting dialogue corpora. One operator uses User Mode to start the dialogue based on the task specification. The other operator takes on the role of Wizard and responds to inquiries in System mode.

APPENDIX C
GENERATED EXAMPLES

A. GENERATED EXAMPLES BY DIFFERENT MODELS

Table 9 shows six examples of generated responses by ToD4IR based on five models. The response is generated based on the given context of the test dataset. Dialogue state and system actions are not shown in this table.

B. WEB INTERFACE FOR ToD4IR

Fig. 10 shows the end-to-end response generation examples by ToD4IR-gpt2-large.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[2] W. Lei, X. Jin, M.-Y. Kan, Z. Ren, X. He, and D. Yin, "Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*. Melbourne, VIC, Australia: Association for Computational Linguistics, Jul. 2018, pp. 1437–1447. [Online]. Available: https://aclanthology.org/P18-1133

[3] S. Akasaki and N. Kaji, "Chat detection in an intelligent assistant: Combining task-oriented and Non-task-oriented spoken dialogue systems," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*. Vancouver, BC, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1308–1319. [Online]. Available: https://aclanthology.org/P17-1120

[4] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E. M. Smith, Y.-L. Boureau, and J. Weston, "Recipes for building an open-domain chatbot," in *Proc. EACL*, 2021, pp. 300–325.

[5] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Jan. 2019, pp. 5370–5381.

[6] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, K. Shuster, E. M. Smith, Y.-L. Boureau, and J. Weston, "Recipes for building an open-domain chatbot," 2020, *arXiv:2004.13637*.

[7] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy, "The pile: An 800GB dataset of diverse text for language modeling," 2021, *arXiv:2101.00027*.

[8] 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," 2020, *arXiv:2010.12757*.

[9] E. Hosseini-Asl, B. Mccann, C. S. Wu, S. Yavuz, and R. Socher, "A simple language model for task-oriented dialogue," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 20179–20191.

[10] 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.* New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 220–224, doi: 10.1145/3434074.3447163.

[11] 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.

[12] 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.*, Nov. 2017, pp. 1–6.

[13] J. Jungbluth, R. Krieger, W. Gerke, and P. W. Plapper, "Combining virtual and robot assistants—A case study about integrating Amazon's Alexa as a voice interface in robotics," in *Proc. Robotix-Acad. Conf. Ind. Robot.*, 2018, p. 5.

[14] Z. Zhang, R. Takanobu, Q. Zhu, M. Huang, and X. Zhu, "Recent advances and challenges in task-oriented dialog systems," *Sci. China Technol. Sci.*, vol. 63, no. 10, pp. 2011–2027, Oct. 2020.

[15] 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*.

[16] 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, doi: 10.1145/3166054.3166058.

[17] 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. 733–743.

[18] J. Perez and F. Liu, "Dialog state tracking, a machine reading approach using memory network," 2016, *arXiv:1606.04052*.

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

[20] 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. 7521–7528.

[21] 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*.

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

[23] 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*.

[24] 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.* New Orleans, LA, USA: Association for Computational Linguistics, Jun. 2018, pp. 2060–2069. [Online]. Available: https://aclanthology.org/N18-1187

[25] B. Dhingra, L. Li, X. Li, J. Gao, Y.-N. Chen, F. Ahmed, and L. Deng, "Towards end-to-end reinforcement learning of dialogue agents for information access," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*. Vancouver, BC, Canada: Association for Computational Linguistics, Jul. 2017, pp. 484–495. [Online]. Available: https://aclanthology.org/P17-1045

[26] T. He, J. Liu, K. Cho, M. Ott, B. Liu, J. Glass, and F. Peng, "Analyzing the forgetting problem in the pretrain-finetuning of dialogue response models," 2019, *arXiv:1910.07117*.
[27] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, and Q. V. Le, "Towards a human-like open-domain chatbot," 2020, *arXiv:2001.09977*.
[28] 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. NAACL-HLT*, 2021, pp. 1570–1583.
[29] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "DIALOGPT: Large-scale generative pre-training for conversational response generation," in *Proc. ACL Syst. Demonstration*, 2020, pp. 270–278.
[30] S. Moon, P. Shah, A. Kumar, and R. Subba, "OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 845–854.
[31] T. Zhao, A. Lu, K. Lee, and M. Eskenazi, "Generative encoder–decoder models for task-oriented spoken dialog systems with chatting capability," in *Proc. 18th Annu. SIGdial Meeting Discourse Dialogue*. Saarbrücken, Germany: Association for Computational Linguistics, Aug. 2017, pp. 27–36. [Online]. Available: https://aclanthology.org/W17-5505
[32] L. Zhou, J. Gao, D. Li, and H.-Y. Shum, "The design and implementation of XiaoIce, an empathetic social chatbot," *Comput. Linguistics*, vol. 46, no. 1, pp. 53–93, Mar. 2020. [Online]. Available: https://aclanthology.org/2020.cl-1.2
[33] Y. Shu, Z. Cao, C. Wang, J. Wang, and M. Long, "Open domain generalization with domain-augmented meta-learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9624–9633.
[34] A. Ghandeharioun, J. H. Shen, N. Jaques, C. Ferguson, N. Jones, A. Lapedriza, and R. W. Picard, "Approximating interactive human evaluation with self-play for open-domain dialog systems," in *Proc. Adv. Neural Inf. Process. Syst., Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., Vancouver, BC, Canada, Dec. 2019, pp. 13658–13669.
[35] P.-E. Mazaré, S. Humeau, M. Raison, and A. Bordes, "Training millions of personalized dialogue agents," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Jan. 2018, pp. 2775–2779.
[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 and L. Lee, "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs," in *Proc. 2nd Workshop Cogn. Modeling Comput. Linguistics*. Portland, OR, USA: Association for Computational Linguistics, Jun. 2011, pp. 76–87. [Online]. Available: https://aclanthology.org/W11-0609
[38] 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. 5016–5026.
[39] 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*.
[40] X. Zang, A. Rastogi, S. Sunkara, R. Gupta, J. Zhang, 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.
[41] E. Dinan, V. Logacheva, V. Malykh, A. Miller, K. Shuster, J. Urbanek, D. Kiela, A. Szlam, I. Serban, R. Lowe, S. Prabhumoye, A. Black, A. Rudnicky, J. Williams, J. Pineau, M. Burtsev, and J. Weston, "The second conversational intelligence challenge (ConvAI2)," Jan. 2019, *arXiv:1902.00098*.
[42] E. M. Smith, M. Williamson, K. Shuster, J. Weston, and Y.-L. Boureau, "Can you put it all together: Evaluating conversational agents' ability to blend skills," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2021–2030.
[43] T. Wolf, V. Sanh, J. Chaumond, and C. Delangue, "Transfertransfo: A transfer learning approach for neural network based conversational agents," *CoRR*, vol. abs/1901.08149, pp. 1–6, Jan. 2019.
[44] R. Lowe, N. Pow, I. Serban, and J. Pineau, "The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems," in *Proc. 16th Annu. Meeting Special Interest Group Discourse Dialogue*. Prague, Czech Republic: Association for Computational Linguistics, Sep. 2015, pp. 285–294. [Online]. Available: https://aclanthology.org/W15-4640
[45] S. Bøgh, M. Hvilshøj, M. Kristiansen, and O. Madsen, "Identifying and evaluating suitable tasks for autonomous industrial mobile manipulators (AIMM)," *Int. J. Adv. Manuf. Technol.*, vol. 61, nos. 5–8, pp. 713–726, Jul. 2012.
[46] 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.
[47] L. El Asri, H. Schulz, S. Sharma, J. Zumer, J. Harris, E. Fine, R. Mehrotra, and K. Suleman, "Frames: A corpus for adding memory to goal-oriented dialogue systems," 2017, *arXiv:1704.00057*.
[48] M. Nardello, O. Madsen, and C. Møller, "The smart production laboratory: A learning factory for industry 4.0 concepts," in *Proc. CEUR Workshop*, vol. 1898, 2017, pp. 1–6.
[49] C. Schou, R. S. Andersen, D. Chrysostomou, S. Bøgh, and O. Madsen, "Skill-based instruction of collaborative robots in industrial settings," *Robot. Comput.-Integr. Manuf.*, vol. 53, pp. 72–80, Oct. 2018, doi: 10.1016/j.rcim.2018.03.008.
[50] C. Fleming, *It's the Way You Say It: Becoming Articulate, Well-Spoken, and Clear*. San Francisco, CA, USA: Berrett-Koehler Publishers, 2013.
[51] 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.
[52] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.
[53] T. Bickmore and J. Cassell, "Small talk and conversational storytelling in embodied conversational interface agents," in *Proc. AAAI Fall Symp. Narrative Intell.*, 1999, pp. 87–92.
[54] T. Klüwer, "'I like your shirt'-dialogue acts for enabling social talk in conversational agents," in *Proc. Int. Workshop on Intelligent Virtual Agents*. Berlin, Germany: Springer, 2011, pp. 14–27.
[55] E. Goffman, *Forms of Talk*. Philadelphia, PA, USA: Univ. of Pennsylvania Press, 1981.
[56] S. Black, G. Leo, P. Wang, C. Leahy, and S. Biderman, "GPT-Neo: Large scale autoregressive language modeling with mesh-TensorFlow," Zenodo, Eur. Org. Nucl. Res., Cern, Switzerland, Tech. Rep., Mar. 2021, doi: 10.5281/zenodo.5297715.
[57] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017.
[58] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy, "The pile: An 800 GB dataset of diverse text for language modeling," 2021, *arXiv:2101.00027*.
[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*, 2001, pp. 311–318.
[60] A. B. Sai, A. K. Mohankumar, and M. M. Khapra, "A survey of evaluation metrics used for NLG systems," *ACM Comput. Surv.*, vol. 55, no. 2, pp. 1–39, Mar. 2023, doi: 10.1145/3485766.

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 the Shanghai Jiao Tong University, Shanghai, in 2015.

From 2015 to 2016, he was a Research Assistant at 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 was a Postdoctoral Researcher at 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.

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 working 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.

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

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 working at the School of Computing and Mathematical Sciences, University of Leicester. He has published 500 papers. His main research interests include knowledge modeling and creative computing. In 2010, he became a Golden Core Member of the IEEE Computer Society.