IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 16, 2019, accepted July 22, 2019, date of publication July 29, 2019, date of current version August 15, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2931576*

# A Novel Slot-Gated Model Combined With a Key Verb Context Feature for Task Request Understanding by Service Robots

**SHUYOU ZHANG, JUNJIE JIANG, ZAIXING HE, XINYUE ZHAO, AND JINHUI FANG**

State Key Laboratory of Fluid Power and Mechatronic Systems, School of Mechanical Engineering, Zhejiang University, Hangzhou 310027, China

Corresponding author: Zaixing He (zaixinghe@zju.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1700504, in part by the National Natural Science Foundation of China under Grant 51775497 and Grant 51775498, and in part by the Natural Science Foundation of Zhejiang Province under Grant LY17F030011.

**ABSTRACT** Spoken language understanding (SLU) is a fundamental to service robot handling of natural language task requests. There are two main basic problems in SLU, namely, intent determination (ID) and slot filling (SF). The slot-gated recurrent neural network joint model for the two tasks has been proven to be superior to the single model, and has achieved the most advanced performance. However, in the context of task requests for home service robots, there exists a phenomenon that the information about a current word is strongly dependent on key verbs in the sentence, and it is difficult to capture this relation well with current methods. In this paper, we extract the key instructional verb containing greater core task information based on dependency parsing, and construct a feature that combines the key verb with its contextual information to solve this problem. To further improve the performance of the slot-gated model, we consider the strong relations between intent and slot. By introducing intent attention vectors into the slot attention vectors through the global-level gate and element-level gate, a novel dual slot-gated mechanism is proposed to explicitly model the complex relations between the results of the ID tasks and SF prediction tasks and optimize the global prediction results. Our experimental results on the ATIS dataset and an extended home service task (SRTR) dataset based on FrameNet show that the proposed method outperforms the most advanced methods in both tasks. Especially, for SRTR, the results of SF, ID, and sentence-level semantic frame-filling are improved by 1.7%, 1.1%, and 1.7%, respectively.

**INDEX TERMS** Human–robot interaction, service robots, slot-gated mechanism, spoken language understanding, verb context feature.

## I. INTRODUCTION

In the past decade, human-computer collaboration promoted by natural language has attracted a great deal of attention in the field of intelligent robots, which includes daily assistance [1], medical care [2], manufacturing [3], indoor or outdoor navigation [4]–[6], and social companionship [7]–[9]. Autonomous service robots are becoming increasingly powerful, and can provide considerable and effective help in the real environment. They need to be able to interact with non-expert users in a natural way, and to understand the high-level task instructions described in natural language in order to plan follow-up tasks to perform low-level actions [10], [11].

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

However, when the robot interacts with the user, the task instructions are usually expressed in natural language that encompasses abstract concepts [12]. In order to execute commands such as "*I want a bowl*," "*Help me with a cup of tea*," and "*Do breakfast*," robots need to know how to map user instructions phrased in an open form to executable operations. This is a key challenge in robot task understanding and has recently attracted a great deal of attention in the artificial intelligence and robotics communities [13].

Many researchers focus on understanding the task intent described in natural language when users interact with service robots. Task intent understanding is translated into intent determination(ID) and slot filling(SF), which is also the goal of this paper. Traditional methods use synonymous verb matching to predict intent. Accuracy is often low given the

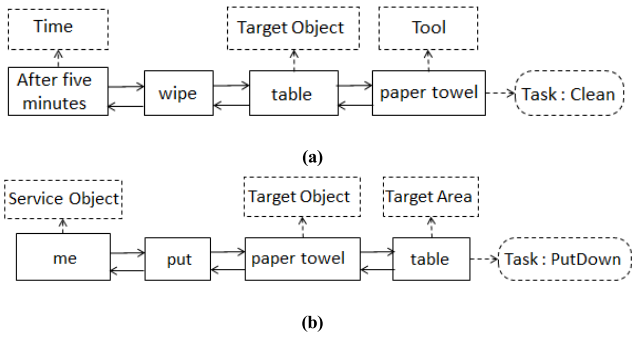



**FIGURE 1.** Information flow in neural networks. (a) The result of the sentence "After five minutes, please wipe the table with a paper towel." (b) The result of the sentence "Please help me put the paper towel on the table."

ambiguity of verbs and incomplete natural language description scenarios [14]. In recent years, with the development of recurrent neural networks, some joint training models based on Long Short-Term Memory(LSTM) and its variants have become widely used in natural language understanding. These models are different from traditional verb matching and other context-free grammar models. They can learn the remote dependencies in sentences through the gating mechanism, and improve the overall performance through joint training of ID and SF. However, this kind of model only adds training loss for the joint training, which cannot model the complex correlation between the two tasks well. Moreover, in the field of service robots, people usually interact with robots in a command language. Unlike auxiliary query languages, many commands in this field contain key verbs that are directly related to behavioral intent. There is a strong information complement relation between many chunks in the command and such key verbs, while there is a weak dependence relation with other chunks. In the example of "*After five minutes, please wipe the table with a paper towel.*" shown in Figure 1 (a), the information of "*Time*", "*Tool*" and "*Target Objec*t" of "*wipe*" are complemented by "*After five minutes*", "*paper towel*" and "*table*". The classification of "*paper towel*" as "*tool*" is largely due to the verb "*wipe*" and has a weak relation with "*five minutes*" or "*table*". In the corpus example "*Please help me put the paper towel on the table.*" shown in Figure 1 (b), the "*paper towel*" acts as "*Target Object*" of "*put*". It can be seen that there is only a weak correlation between the "*paper towel*" and words such as "*table*", but a strong correlation between the "*paper towel*" and the key verbs such as "*wipe*" and "*put*" in the command.

The focus of this paper is to propose a method to fully mine the association information between key verbs and context words in task requests, and to create a mapping from sentence to expected robot action frame. We have noticed that task-oriented oral descriptions usually contain multiple verbs with redundant information and a key verb that is directly related to the task. We propose a method of extracting the key verb

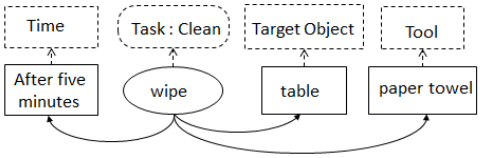



**FIGURE 2.** Information flow that our approach encourages.

based on dependency parsing and a way of combining the key verb with its contextual information to improve the performance in task understanding for service robots. As shown in Figure 1 (a), related works attempt to let the neural network automatically learn how strong the task intents depend on each word in the text. As shown in Figure 2, we construct a new feature to input the priori dependence between words and key verbs into the neural network, and encourage the network to learn the strong dependence of the slots on key verbs, which is very helpful for task request understanding in the field of service robots. In addition, we propose a novel dual slot-gated mechanism, which includes a global-level gate and an element-level gate to explicitly model the complex relations between ID and SF tasks, so as to control intent vectors to guide the prediction of slot context vectors. By introducing the features of key verbs and their contextual information into the model input, the performance of our proposed recurrent neural network with a slot-gated mechanism in the joint ID and SF tasks is improved. The proposed structure is applied to the ATIS dataset and a domain corpus of service robots constructed by the author's laboratory based on the FrameNet dataset.

The contributions are three-fold: 1) Aimed at natural language task requests for service robots, a feature that combines key verbs and their contextual information is used in the input of the proposed neural network model, which makes good use of the core information of instructional verbs in home service-oriented task description. 2) Compared with the most advanced neural network model and traditional key verb matching model, the proposed dual slot-gated neural network model achieves better performance by explicitly modeling complex intent-slot relations through the global-level gate and the element-level gate. 3) We test our model on the ATIS dataset and an extended home service task dataset based on FrameNet, and the experiments show that our model performs better on the two datasets, which proves the robustness and effectiveness of the proposed model.

This paper is organized as follows: Section 2 summarizes the earlier related work and its limitations. The new feature and the corresponding general extraction method are introduced in Section 3, and then the structure of the novel dual slot-gated mechanism model is described in detail. Details of the experiment and analysis of the results are presented in Section 4. Finally, the concluding remarks and directions for future work are discussed in Section 5.

## II. RELATED WORKS

Usually, task request in the form of natural language are open and complex, and it is difficult for robots to understand instructions that are not expressed with sufficient clarity. Therefore, robots first need to parse the semantics of the user tasks and translate user tasks into an internal representation that the machine can understand. This task encounters two main challenges. The first one is the semantic understanding of ambiguous natural language. For example, an instructional verb can have multiple meanings. The second one is to map the content of the language to the real object in the robot's running environment, i.e. grounding processing. This paper focuses on the first type of challenges. In this challenge, many practical task-oriented conversation systems have been developed for robot task understanding. The purpose of these systems is to automatically detect the user's intent expressed in natural language as the basis for subsequent task planning. Appropriate actions to be taken by the system are determined by a defined intent. So far, there are many methods in this regard that can be used for robot task understanding. Most of the research focuses on building a frame template for high-level tasks, and filling and instantiating the frame template with semantic roles in user task description. More precisely, user instructions or high-level tasks are parsed into a series of frame elements based on predefined knowledge. These frame templates come from open resources [14] or domain experts [15], for example, the words "*drink*" and "*fridge*" in the user task "*Serve a drink from fridge*" can be filled into the elements "*Theme*" and "*Source*" of the frame "*Bringing*," respectively. Although this task has been extensively studied, intent understanding is still not considered to be a problem that is solved. This is because when the system has to deal with spoken language task requests with ambiguous and incomplete information, automatic mapping from natural language to the frame template and automatic filling of the template elements become challenging [16].

In the traditional method, the keyword matching method uses domain dictionary to parses instructions, and the matching results are used as trigger conditions of the frame template [17]–[19]. Unfortunately, there are two main problems with this approach. First, natural language is ambiguous, and the same verb may cover multiple tasks. For example, when a user's task is described as "*I want to drink water*," the command means to fetch water, that is, to execute the command related to "*Bringing*," rather than let the robot execute the command "*Drinking*." Second, spoken language descriptions sometimes omit prepositions or other components, which makes task intent matching difficult. Some studies introduce a synonym dictionary [20], such as WordNet, to parse words with similar semantic similarity into the same object, but there are still a large number of action instructions that cannot be matched correctly. Some studies attempt to use context-free grammar models to parse task requests [21], [22], but the reason these grammar models are ineffective in practice may be that such grammar models have difficulty in modeling long-distance dependency phenomena. In addition, these studies can only parse a small number of semantic roles by learning simple grammatical rules, which is different from the process of filling in frames that contain a large number of semantic roles in this paper. Introducing a user clarification mechanism to enable robots to acquire the missing key information through dialogue and inquiry is also a method to improve task understanding performance [23]. However, friendly human-computer natural language interaction systems should have a certain degree of reasoning ability to deal with missing or incomplete natural language instructions [24].

In recent years, scholars have noticed that when human beings give instructions, a pattern exists such that key instructions can usually be inferred from the overall task description [25]. Some studies in the field of spoken language understanding use statistical learning methods, which regards ID as the classification of question semantics, and use the SF method to parse semantics. Some studies have shown that the neural network model has significantly improved the accuracy of general models. Popular methods include recurrent neural network (RNN) [26], the joint model for ID and SF [27], and the attention-based model [28]. Most of these studies focus on simple query statements, but the performance improvement in task understanding for service robots is not significant [29]. The reason may be that common query task descriptions, such as "*What are the flights from Tacoma to San jose*," differ from service language which usually contains verbs carrying core instructional information, such as "*carry*" in "*Could you please help me carry this book to the bedroom*." In addition, the joint training model proposed in recent years "implicitly" combines the ID and SF tasks, and cannot adequately extract effective information from the ID tasks to guide the SF tasks. Recently, the slot-gated mechanism has been proposed to guide joint optimization by simply modeling the overall proportional relationship between the two tasks [30]. However, the simple method of constructing a trainable global weight for intent vectors and slot vectors cannot model complex intent-slot correlation that well.

## III. PROPOSED APPROACH

In this section, we explain the method proposed for SLU in the service robots domain, which is used to explain the sequence of instructions for robot tasks expressed by non-expert users in natural language. First, the key verb feature and its extraction method proposed for model input is introduced. Then, our attention-based RNN model is explained. Finally, the joint training model based on our slot-gated mechanism is introduced.

### A. PROBLEM DEFINITION

We translate task request understanding by service robot into ID and SF problems. Given sentence $X$, the ID task is to classify $X$ into frame $f_k$ in frame set $F = \{f_1, \ldots, f_{|F|}\}$, and the SF task is to select slots in frame $f_k$ to be filled with lexical chunks in $X$. For SF tasks, specifically, we assume

that $\Re_f = \left\{r_1, \ldots, r_{|\Re_f|}\right\}$ represents all slots belonging to frame $f_k$, and $S$ is a candidate continuous chunk in a sentence. If there are slots that are not explicitly filled by chunks, they are filled by empty strings. Assuming that $A$ represents the mapping of slot $\mathcal{R}_f$ to chunk $S$, the SF method used in the past is to predict each slot $r_{k \in} \mathcal{R}_f$:

$$A\left(r_k\right) = \arg\max_{s \in S} p\left(s | r_k, f_k, X\right) p\left(f_k | F, X\right) \tag{1}$$

Now we propose a new method. We define the concept of key verb $v$ and propose a new feature, which uses the information of dependency context $T$ and word context $v_c$ to mark slots into chunks related to key verbs. Then (1) is reformed as (2):

$$A\left(r_k\right) = \arg\max_{s \in S} p\left(s | r_k, f_k, T, v_c, X\right) p\left(f_k | F, T, v_c, X\right) \tag{2}$$

### B. OVERALL ALGORITHM AND METRICS

As illustrated in Figure 3(a), the proposed architecture is composed of five components: feature processing, Bi-LSTM layer, attention layer, dual slot gated mechanism.

The evaluation metrics used for labeling include accuracy, precision, recall, and F1-score. Their calculation formulas are as follows:

$$Accuracy = \frac{TP + TN}{N} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = \frac{2TP}{2TP + FN + FP} = \frac{2 * precision * recall}{precision + recall} \tag{6}$$

Among them, the prediction is a positive example, the actual is also a positive example, we call it true positive (TP), the second case, the prediction is a positive example, the actual is a negative example, we call it false positive (FP), the third case The prediction is a negative example, which is actually a positive example, called false negative (FN). In the last case, the prediction is a negative example, and the actual is also a negative example, called true negative (TN).

### C. FEATURE BASED ON KEY VERB CONTEXT

Some studies have shown that adding hand-crafted features to the input of neural networks can improve prediction performance [31]. Different from common query tasks, task requests in service robot domain have characteristics such that the core information of a task's overall intent is often contained in some key command verbs, while other verbs contain redundant information to assist understanding.

We define the concept of key verbs. Key verbs are a class of verbs that contain core instruction information in task requests. They are usually uniquely identified in a single instruction. In the related work, scholars try to let the three

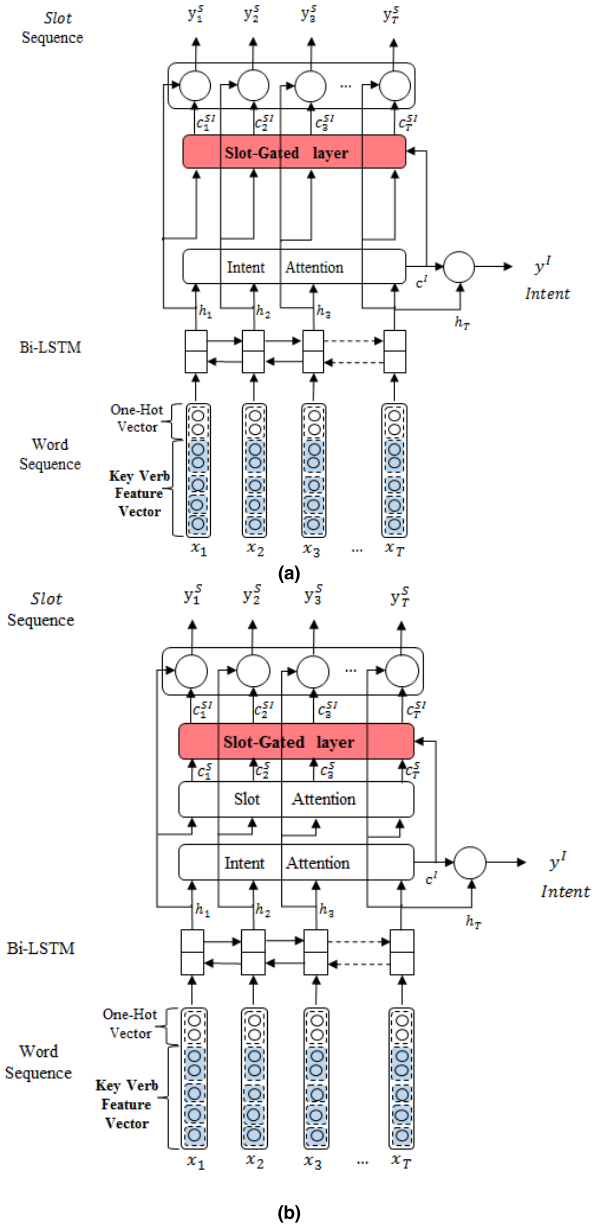



**FIGURE 3.** The structure of the proposed model. (a) slot-gated model with intent attention. (b) Slot-Gated model with full attention.

gated mechanisms of the LSTM neural network model automatically learn the dependence of the prediction of the current word on each word in the sentence, but they cannot make good use of the strong dependence of the current word on such key verbs. Therefore, based on the dependency path of verbs, we explicitly define the extraction method of key verbs. However, given the influence of different users' spoken expressions, the intent of the key verb is often ambiguous and needs to be enriched by its contextual information. Based on such characteristics, we propose a new feature, which is based on the key verbs and their contextual information, it can effectively use the correlation information between each word in the input sequence and the context of the key verbs to

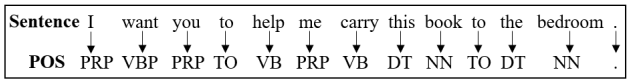



**FIGURE 4.** An example utterance with annotations from the Stanford Part of Speech Tagging.

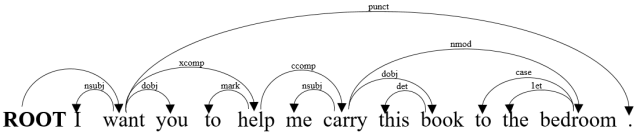



**FIGURE 5.** An example utterance with annotations of stanford dependent syntax annotation.

guide the final prediction results. We process the sequence word for word. In this feature, there are four key parameters: the key verb $x_{key}$, the key verb context $x_{context}$ and the region mark $x_{mark}$, the dependency context $x^D_{context}$.

The first step is to extract the key verb of the input sentence $x_{sen} = \{x_1, x_2, \ldots, x_T\}, t = 1, 2, 3 \ldots, T$. Part-of-speech (POS) tagging and dependency parsing are basic modules in natural language processing. At present, many effective and general parsers, such as Stanford Parser, have been used by scholars to lay a foundation for subsequent information extraction. The POS of a word can be divided into verbs, nouns, adjectives and so on according to its meaning, form and grammatical function in the language it belongs to. Using the Stanford parser, when the input word sequence is "*I want you to help me carry this book to the bedroom*," the result of POS tagging is shown in Figure 4. Then the verbs contained in this sequence are extracted as follows: "*want*," "*help*," "*carry*."

The task of dependency parsing is to analyze the dependency relations between words in a given sentence [32]. With the same parser, when input is the natural language instruction mentioned above, the result of dependency parsing is shown in Figure 5.

Mark the dependency path depth of each verb to the root node in the sentence as $n_{depth} = \{n_1, n_2, n_3, \ldots, n_T\}$ , that is, the number of arcs contained between the root node and the selected verb in Figure 5. Then the dependency paths of the verbs "*want*," "*help*" and "*carry*" are as follows: $n_{want} = 1$, $n_{help} = 2, n_{carry} = 3$.

We have noticed that redundant information often exists in spoken language, and a few of the extracted verbs contain the core intent information of the instructions and most of them contain auxiliary information. The deeper the dependency path is, the more core intent information the verbs contain. Therefore, the following rules are formulated:

If the maximum element $n_{max}$ in $n_{depth}$ is unique, $x_{key}$ is the word vector corresponding to $n_{max}$, otherwise $x_{key}$ is equal to the sum of the corresponding word vectors. If the result of POS tagging does not contain verbs, $x_{key} = 0$.

Then $x_{key} = x_{carry}$ in the sequence mentioned above.

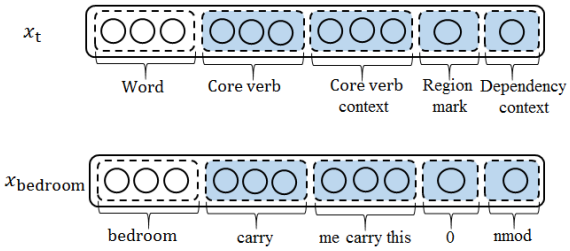



**FIGURE 6.** The input vector for each word. Shadow part is the new constructed key verb context vector.

Because just one key verb cannot accurately describe the action information, we set up a context window with the length n∈N to construct the key verb feature. By extending the context and using the key verb context $x_{context}$, ambiguity can be largely eliminated . Context window is a common parameter in semantic role labeling task. Because of the similarity between two tasks, we use the settings in reference [31] and set it to 3 in our model., that is to say, $x_{context}$ is the vector addition of the key verb and the words before and after. If a word is in the context window of a key verb, we use the region mark $x_{mark} = 1$ to denote the word position, otherwise use $x_{mark} = 0$ . In addition, we construct the dependency context of key verbs $x^D_{context}$. This vector is the type of dependency between the current word and the extracted key verb. If there is no dependency, it is 0 vector. The input vectors of the model are obtained by concatenating the four feature vectors with the original word vectors, as shown in Figure 6.

### D. ATTENTION-BASED RNN MODEL

The Long Short-Term Memory (LSTM) model is a special RNN model. It can learn long-term dependency and has strong expressive ability for natural language [33]. Generally, LSTM units are composed of three gated organizations, which are used to control the proportion of information that needs to be forgotten in the one-way transmission. Figure 7 shows the general LSTM network structure. An LSTM unit maps a word vector $x^{<t>}$ to a one-way hidden state $a^{<t>}$.

Formally, the formula for updating LSTM cells at $t$ time is as follows:

$$O^{<t>} = \tanh\left(W_o\left[a^{<t-1>}, x^{<t>}\right] + b_o\right), \quad (7)$$

$$\delta_{update} = \sigma\left(W_{update}\left[a^{<t-1>}, x^{<t>}\right] + b_{update}\right), \quad (8)$$

$$\delta_{forget} = \sigma\left(W_{forget}\left[a^{<t-1>}, x^{<t>}\right] + b_{forget}\right), \quad (9)$$

$$\delta_{output} = \sigma\left(W_{output}\left[a^{<t-1>}, x^{<t>}\right] + b_{output}\right), \quad (10)$$

$$O^{<t>} = \delta_{update} \cdot \widetilde{O}^{<t>} + \delta_{update} \cdot O^{<t-1>}, \quad (11)$$

$$a^{<t>} = \delta_{output} \cdot \tanh O^{<t>}, \quad (12)$$

where *tanh* and $\sigma$ are the activation functions, and each gate $\delta$ contains a trainable weight matrix $W$ and bias $b$.

Figure 8 presents a Bi-LSTM structure, which is proposed as an improved LSTM structure to improve the disadvantage

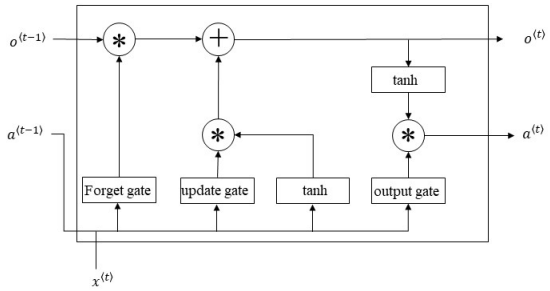



**FIGURE 7.** The long short-term memory cell.

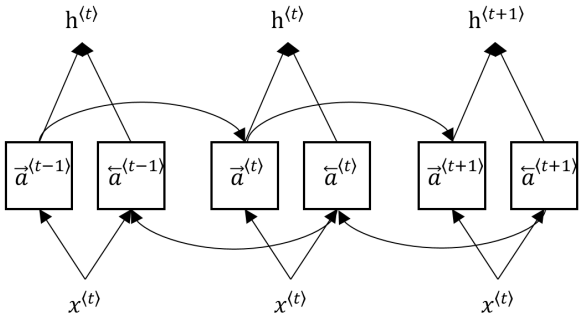


**FIGURE 8.** A bidirectional LSTM network.

in which LSTM can not access the future information as input to affect the current timing prediction [34]. Its basic idea is to present each sequence forward and backward as two separate hidden states to capture the past and future information respectively, and then connect the two hidden states to form the final output $h^{<t>}$:

$$h^{<t>} = \tanh(W_h \left[ \overrightarrow{a}^{<t>}, \overleftarrow{a}^{<t>} \right] + b_h), \tag{13}$$

where *tanh* is the activation function, $W_h$ is the weight matrix, and $b_h$ is the bias.

The attention mechanism breaks the restriction in the traditional encoder-decoder structure that is a dependency on a fixed length vector in encoding and decoding process [35]. It retains the intermediate output of the input sequence from the LSTM encoder, and then trains a model to selectively learn these inputs and associates the output sequence with the model output. In other words, the generating probability of each item in the output sequence depends on which item is focused on in the input sequence.

For the SF task, attention weights $\alpha^S_{i,j}$ and slot labels $y^S_i$ can be calculated using the following formula:

$$c^S_i = \sum_{j=1}^{T} \alpha^S_{i,j} h_j, \tag{14}$$

$$\alpha^S_{i,j} = \frac{\exp\left(e_{i,j}\right)}{\sum_{k=1}^{T} \exp\left(e_{i,k}\right)}, \tag{15}$$

$$e_{i,k} = \sigma\left(W^S_{he} h_k\right), \tag{16}$$

$$y^S_i = soft\ \max\left(w^S_{hy}\left(h_i + c^S_i\right)\right), \tag{17}$$

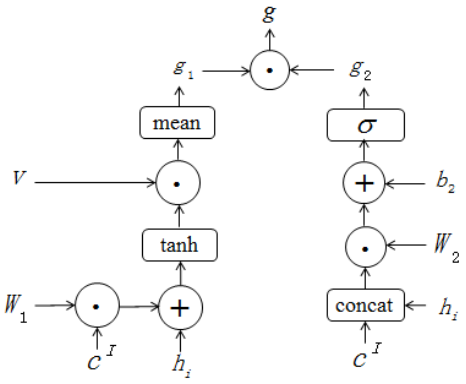



**FIGURE 9.** Illustration of the dual slot gate.

where $y^S_i$ is the slot label of the i-th word in the input, and $w^S_{hy}$ is the weight matrix.

For the ID task, intent labels can be calculated using similar formulas:

$$c^I = \sum_{i=1}^{T} \alpha^I_i h_i, \tag{18}$$

$$\alpha^I_i = \frac{\exp\left(e^I_i\right)}{\sum_{k=1}^{T} \exp\left(e^I_k\right)}, \tag{19}$$

$$e^I_k = \sigma\left(W^I_{he} h_k\right), \tag{20}$$

$$y^I = soft\ \max\left(w^I_{hy}\left(h_T + c^I\right)\right), \tag{21}$$

### E. JOINT OPTIMIZATION MODEL WITH DUAL SLOT-GATED MECHANISM

This section describes the joint optimization model for the SF and ID tasks. Based on the idea that ID and SF tasks are related, a joint optimization model is proposed to improve the prediction performance of single task model, where the model only achieves the goal of joint optimization by simply adding the loss of the two task training models. This method implicitly models the correlation between ID results and SF results, and tries to let the neural network automatically learn the proportional contribution of each of the two tasks to sentence-level semantic annotation. In this section, we propose a novel dual slot-gated mechanism, which controls the intent attention vector to predict the results of the slot attention vector through two gates, i.e. the global-level gate and the element-level gate, so as to optimize the results globally. The proposed model can improve the performance of the attention-based joint optimization model by explicitly constructing the contribution ratio of ID task results to SF task prediction. Our model is different from the slot-gated model proposed by Goo et al., which uses only a single gate to model the global contribution ratio of ID task to SF task [30].

The red annotated part in Figure 3 is the proposed slot-gated layer. As shown in Figure 9, our dual slot-gated mechanism introduces two additional gates, i.e. the global-level gate $g_1$ and the element-level gate $g_2$ , which model slot-intent

relations by using an intent context vector to influence the slot context vectors prediction. First, the slot context vector is combined with the intent context vector $c^I$ , then activated and weighted by a slot gate. The vector obtained is averaged on the vector dimension and mapped to a variable $g_1$:

$$g_1 = \frac{1}{n}\sum v \cdot \tanh\left(h_i + W_1 c^I\right), \tag{22}$$

where $v$ and $W_1$ are the trainable vector and matrix respectively, $n$ is the dimension size of $c_i^S$. $g_1$ can be regarded as a weighted global feature, which controls the overall contribution ratio of the slot context vector and intent vector to slot label prediction. A larger $g_1$ indicates that the correlation between tasks is greater, and the prediction of this slot label is more dependent on the result of intent vector prediction.

In order to get the contribution ratio of each element of the intent context vector to the slot label prediction, we use the activation function $\sigma$ to process the joint vector:

$$g_2 = \sigma\left(W_2\left[h_i, c^I\right] + b_2\right), \tag{23}$$

The gate $g_2$ filters out the elements of the intent vector that are beneficial to SF tasks and assigns them a ratio of 0 to 1. Multiply $g_1$ and $g_2$ to obtain a $g$ vector, which contains the overall correlation ratio and element-level correlation ratio between the two task vectors:

$$g = g_1 \cdot g_2, \tag{24}$$

Then we apply g to the joint context vector to get the final slot context vector $c_i^{SI}$ and replace (11) as below:

$$c_i^{SI} = g \cdot \tanh\left(h_i + W_3 c^I\right), \tag{25}$$

$$y_i^S = soft\max\left(W_{hy}^S\left(h_i + c_i^{SI}\right)\right), \tag{26}$$

In general, applying the attention layer only to ID tasks will achieve better results [30]. In order to analyze the combined performance of the slot-gated mechanism and the attention mechanism, we also propose a slot-gated model using full attention as shown in Figure 3(b). Then (22), (23) and (25) are reformed as (27), (28) and (29), respectively.

$$g_1 = \frac{1}{n}\sum v \cdot \tanh\left(c_i^S + W_1 c^I\right), \tag{27}$$

$$g_2 = \sigma\left(W_2\left[c_i^S, c^I\right] + b_2\right), \tag{28}$$

$$c_i^{SI} = g \cdot \tanh\left(c_i^S + W_3 c^I\right), \tag{29}$$

The target of joint prediction of SF and ID is expressed as

$$\begin{aligned} &p\left(y^S, y^I|X\right) \\ &= P\left(y^I|X\right)\prod_{t=1}^{T} P\left(y_t^S|X\right) \\ &= P\left(y^I|x_{1,\ldots,}x_T\right)\prod_{t=1}^{T} P\left(y_t^S|x_{1,\ldots,}x_T\right), \end{aligned} \tag{30}$$

where $p\left(y^S, y^I|X\right)$ is the conditional probability of the SLU result given the input word sequence and need to be maximized.

**TABLE 1. Pre-programmed robot actions "*bringing*" with associated parameters, with abbreviations in parentheses.**

| Semantic roles in the frame "Bringing" | Abbreviation |
|---|---|
| Agent | age |
| Service Object | so |
| Target Object | to |
| Target Area | ta |
| Way Point | wp |
| Path End Point | pep |
| Time | tim |
| Tool | too |
| Speed | spe |
| Explanation | exp |
| Purpose | pur |
| Frequency | fre |

## IV. EXPERIMENT

In order to evaluate the robustness of the proposed model, we conducted experiment 1 on a benchmark dataset, the Air Travel Information System (ATIS) dataset. In addition, we extracted the frame of instructional verbs related to service robots in the FrameNet dataset, and collect a new corpus(SRTR) by manual annotation. Experiments 2 are conducted on this corpus to verify the enhanced effect of our model and new features on the understanding of spoken language instructions in the service robot domain.

### A. DATASET

The ATIS dataset [36], [37], which consists of recordings of flight reservation personnel, is widely used in the field of spoken language understanding. The training set contains 4978 questions from the ATIS-2 and ATIS-3 corpus, and the test set contains 893 questions from NOV93 and DEC94 in ATIS-3. The dataset has 127 different slot labels and 18 different intention types.

In addition, we collected a corpus in the field of service robots task requests based on the FrameNet dataset by manual annotation. FrameNet [38], [39] is a lexical resource based on frame semantics constructed by the University of California, Berkeley, USA. It contains abundant semantic information and a large number of English sentence analysis examples. In FrameNet, "frame" is a linguistic term used to describe a set of concepts of an event or a semantic scene. Each frame contains a series of semantic roles called *frame elements*, which correspond to words describing events or event forms in context. We extract and extend the instructional verb frames related to the service robot domain, so that the semantic roles of the frame template can more fully cover the natural language instruction in home service domain. The candidate semantic roles used for expansion include: "*Agent*," "*Service Object*," "*Target Object*," "*Target Area*," "*Way Point*," "*Path End Point*," "*Time*," and "*Tool*." The table 1 shows the content of the frame "*Bringing*."

This paper simulates three family environments, living room, kitchen and bedroom, and extracts 12 frames related to service instructions in FrameNet, including: "*Bringing*," "*Cutting*," "*Filling*," "*Getting*," "*Giving*," "*Grasp*," "*Grinding*," "*Placing*," "*Scouring*," "*Chatting*," "*Installing*," and "*Removing*." Then free format natural language instructions and robotic instructions logs are collected from multiple users and labeled manually with BIO tags. For the ID task, we mark the entire sentence with the frame it belongs to. For the SF task, we mark each word in the sentence with its own slot. If a word is the beginning of a slot, it is marked as "*B-label"*, if it is inside the slot but not the first word, it is marked as "*I-label*," otherwise it is marked as "*O*". In the living room environment shown in Figure 10, an example of natural language instructions and corresponding annotations collected under the instruction frame "*Bringing*" are shown in Figure 11. In the example sentence "*Robot 2 fetch me the coffee from the table in five minutes*," the phrases "*Robot 2*," "*me*," "*coffee*," "*table*," and "*in five minutes*" are respectively marked as the semantic roles "*Agent*," "*Service Object*," "*Target Object*," "*Target Area*" and "*Time*" under the frame "*Bringing*."

The SRTR dataset was manually annotated and tested by 11 workers in the laboratory where the author works. In SRTR, we collect the description texts by the following method: We invited 40 participants aged between 20 and 60 from various industries to participate and show them the constructed family environment and the list of candidate frames. In order to encourage diversity, we required each assistant to provide more than 10 task requests for each framework, so the collected request descriptions cover all tasks in a more balanced manner. We construct a set of pipelines for annotation. The annotation process is divided into the following stages: We construct a set of pipelines for annotation. The annotation process is divided into several stages as follows: 1. Given a description of a task request, one annotator is randomly assigned, and the annotator chooses to submit a annotation or a deletion application. The reason for deletion is usually because the task request contains a description that cannot be uniquely determined. 2. After the initial annotation is completed, it enters the verification stage, which adopts the majority voting strategy. Three workers are randomly assigned to each annotation to verify, and vote on whether it is correct or not, or vote on deletion application. The annotation will only be added to SRTR if at least two workers think it is correct. The controversial description will return to the first step for iteration. 3. After 5 iterations, the remaining descriptions are deleted and the SRTR data set is obtained.

The training set contains 2808 task requests from the corpus and the test set contains 312 task requests from the corpus. The data set has 61 different slot labels and 12 different intent types. Compared with the ATIS dataset, the SRTR dataset is more oriented toward complex task execution than simple task queries. According to statistics, there are no more than four frames covered by the same key verb in SRTR. In addition, intents in ATIS are highly unbalanced, where "*atis_flights*" accounts for about 74% of the training data and "*atis_cheapest*" appears only once [30]. The corpus of all types of intent in the SRTR dataset is kept in balance basically.

**TABLE 2. Hyperparameter settings in our experiments.**

| Hyperparameter | Value |
|---|---|
| Batch Size | 16 |
| Patience | 5 |
| Clip Norm | 5 |
| Input Dropout | 0.5 |
| Output Dropout | 0.5 |
| Window length | 3 |
| Learning Rate | 0.001 |
| Beta1 | 0.9 |
| Beta2 | 0.999 |
| Epsilon | 10E−8 |

### B. SETUP

In experiment 1, our input is one-hot word vector, and in experiment 2, our input is a combination of a one-hot word vector and the new feature vector proposed in section 3. We set the size of the hidden vector to 64, the optimizer to Adam, the reported number averaged in 30 runs, and set the maximum epochs to 50 and 40 on ATIS and SRTR respectively by using the early stop strategy. In experiment 2, the traditional method for the ID task was added to the baseline, which mapped the natural language instructions containing verbs with similar meanings to the same frame using the WordNet synonym dictionary.

In order to alleviate the gradient explosion phenomenon, we used gradient clipping in the training process and set clip norm to 5. To prevent overfitting, dropout rate of input and output layers is set to 0.5 in Bi-LSTM model. The main super-parameters that have been fine-tuned include batchsize, context window length, learning rate, and the exponential decay method is used to adjust the learning rate in the later stage of training to achieve higher accuracy. The specific hyperparameter settings are shown in Table 2.

### C. RESULTS AND ANALYSIS

In this paper, the F1 score is used to evaluate the performance of the SF task, accuracy is used to evaluate the performance of the ID task, and full frame accuracy is used to evaluate the performance of sentence-level semantic frame analysis.

The results of experiment 1 are shown in the table 3, in which the comparison baselines for SF and ID include the state-of-the-art methods using the Bi-LSTM [40] model, Attention-Based model [28], and Slot-Gated model with a single gate [30].

Table 3 shows that the model with the slot-gated mechanism achieves significantly better performance than the baseline model in all tasks (slot filling, intention prediction and

**TABLE 3.** SLU performance on the ATIS dataset compared with previous approaches.

| Model | | ATIS Dataset | | |
|---|---|---|---|---|
| | | Slot (F1) | Intent (Acc) | Sentence (Acc) |
| Bi-LSTM | | 94.3 | 92.6 | 80.7 |
| Attention-Based | | 94.2 | 91.1 | 78.9 |
| Slot-Gated (Single Gate) | Full Attention | 94.8 | 93.6 | 82.2 |
| | Intent Attention | 95.2 | 94.1 | 82.6 |
| Proposed Slot-Gated | Full Attention | 95.1 | 93.8 | 82.4 |
| | Intent Attention | 95.5 | 94.3 | 83.0 |

**TABLE 4.** SLU performance on the SRTR dataset compared with previous approaches.

| Model | | SRTR Dataset | | |
|---|---|---|---|---|
| | | Slot (F1) | Intent (Acc) | Sentence (Acc) |
| Verb Matching | | - | 73.5 | - |
| Attention-Based | | 91.2 | 87.8 | 76.9 |
| Slot-Gated (Single Gate +Intent Attention) | | 91.7 | 88.2 | 77.7 |
| Proposed | Slot-Gated(Intent Attention) | 92.3 | 88.6 | 78.2 |
| | Slot-Gated(Intent Attention)+ Key Verb Context Feature | 93.4 | 89.3 | 79.4 |

**TABLE 5.** Ablation studies on the SRTR dataset.

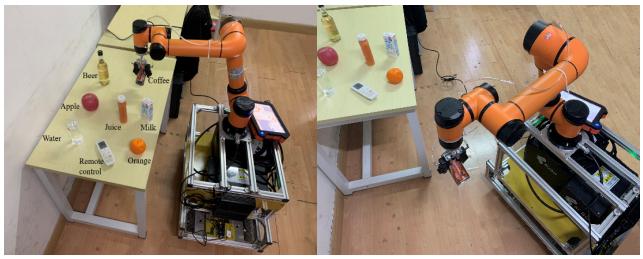
| Model | SRTR Dataset | | |
|---|---|---|---|
| | Slot (F1) | Intent (Acc) | Sentence (Acc) |
| Word Context Feature | 93.1 | 89.2 | 78.9 |
| Dependency Context | 92.7 | 88.8 | 78.7 |
| Key Verb Context Feature (fully) | 93.4 | 89.3 | 79.4 |

semantic frame). In particular, the result of the sentence-level semantic frame has been greatly improved, which indicates that the slot-gated mechanism provides beneficial information for the global optimization of the joint model by learning the correlation between intent and slot. Compared with the original method of simply using a single gate to model the overall contribution ratio between tasks, our slot-gated model has made further improvements in the results of all tasks. The reason is that our model considers the contribution ratio of intent vector to slot vector prediction task at the element level and the global level, so that our model can learn more complex inter-task dependencies to optimize the joint task. In addition, by analyzing the experimental results of slot-gated models with different attention mechanisms, we note that the results of the model with intent attention are better than those of the model with full attention. Considering that slot annotation of current words relies more on the overall intent prediction results than the slot information of long-distance words in sentences, the possible reason is that our slot-gated layers can model the dependencies between tasks well, without the simple weighting operation of slot vectors by slot-attention layer.

The results of experiment 2 are shown in Table 4. The baseline of comparison includes the traditional verb matching method based on WordNet synonym dictionary, the attention-based Bi-LSTM model and the slot-gated model with single gate and intent attention. The results show that the accuracy of the ID task is greatly improved by using the neural network model based on joint optimization. One of the possible reasons for this improvement is that some instructions issued in spoken language do not contain retrievable verbs in synonym dictionaries, which results in matching failure, and the use of the recurrent neural network model can infer the intents of instructions from contextual information. In this dataset, our slot-gated layer also appreciably improves the results by adding an additional element-level gate to explicitly model more complex correlation of slots and intents, and the relative improvements are 0.5%, 0.4%, 0.5%, respectively, which proves the effectiveness of the proposed model. One possible reason for the greater improvement compared with the experimental results on the ATIS dataset is that the complexity of the data set is different. For a relatively simple SLU task, such as understanding of query statements in ATIS, the joint training model can achieve better results by adjusting the overall contribution ratio of each task reasonably, without adding additional gates to model the element-level contribution ratio. In addition, SF, ID and semantic frame tasks were significantly improved after adding new features to our model, especially the SF task, which increased by 1.1%. The possible reason is that the information from the key verb and its contexts plays an important role in guiding semantic role tagging in different semantic frames. Experiments show that the proposed model can effectively improve the performance of intent understanding of task requests for home services by utilizing the key verbs and their contextual information. Overall, the experimental results of the proposed method are 1.7%, 1.1% and 1.7% higher than those of the baseline method, respectively.

We conducted a ablation study to verify the validity of the proposed modules in the context features of key verbs, i.e. word context feature and dependency context feature. The results of the ablation study are shown in Table 5. We remove each module to verify the effectiveness of using all proposed modules. When we use all the modules together, we can clearly see the performance improvement. The results show that both modules play a key role. Using the word context feature is critical, and using the dependency context feature further improves performance.

Interestingly, our model performs better than the baseline method in filling in slots specific to the frame. For example, when testing on the sentence "*Help me find the lost wallet,*




(a) (b)

**FIGURE 10.** Service robot response to a natural language task request in a living room environment: (a) shows that the service robot is choosing to grab the coffee on the table; and (b) shows that the service robot is delivering the coffee to the end point.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| W | Robot | 2 | fetch | me | the | coffee | from | the | table | in | five | minutes |
| S | B-age | I-age | O | B-ser | O | B-to | O | O | B-ta | B-tim | I-tim | I-tim |
| I | Bringing | | | | | | | | | | | | |

**FIGURE 11.** An example utterance with annotations of semantic slots in the BIO format (S) and intent (I).

*possibly on the table or on the way to the sofa.*", our method can correctly mark "*table*" as the slot "*Path Start Point*" under the frame "*Scouring*", and the baseline model will mistakenly identify it as "*Target Area*", which is common in most frames. The possible reason is that introducing the information of key verbs can help the model to extract the dependence relationship between slots and frames, and the contribution proportion of task intents covered by key verbs to semantic role recognition is explicitly learned through the dual slot gated mechanism.

The experiment runs on a single NVIDIA 1080Ti GPU server. The offline training and online testing methods are used and the online testing phase is completed within 0.5s. The average processing time of a single task request is 1ms, which basically meets the requirement of real-time human-robot interaction.

## V. CONCLUSION AND FUTURE WORK

Using the characteristics of spoken task requests in the field of home service robots, this paper proposes a task understanding method based on the key verb and its contextual features. The model with this new feature is more effective than the traditional method of using only one-hot word vectors, which shows that adding the proposed feature can help the model extract the core information from verbs used in home service instructions for task understanding. In addition, we add a novel dual slot-gated layer to the latest attention model to learn the complex intent-slot relations explicitly, so that SF can be conditioned by the results of intent learning. Applying this model to the ATIS and home service datasets, we obtain better results than the baseline, which shows that the proposed model achieves better SLU (joint slot filling and intent determination) performance. It also shows that the model has robustness in SLU tasks in different domains.

In future work, we plan to propose an architecture that can better show the correlation between ID and SF. Since the gated structure of this paper is one-way, we will study further the structure of the explicit relations that input SF results into ID or provides a bidirectional guide.

## REFERENCES

[1] N. Ayari, A. Chibani, Y. Amirat, and E. Matson, "A semantic approach for enhancing assistive services in ubiquitous robotics," *Robot. Auton. Syst.*, vol. 75, pp. 17–27, Jan. 2016.
[2] M. Gnjatović, "Therapist-centered design of a robot's dialogue behavior," *Cogn. Comput.*, vol. 6, no. 4, pp. 775–778, Dec. 2014.
[3] R. Liu and X. Zhang, "Generating machine-executable plans from end-user's natural-language instructions," *Knowl.-Based Syst.*, vol. 140, pp. 15–26, Jan. 2018.
[4] Z. Huo, T. Alexenko, and M. Skubic, "Using spatial language to drive a robot for an indoor environment fetch task," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Chicago, IL, USA, Sep. 2014, pp. 1361–1366.
[5] C. Landsiedel, V. Rieser, M. Walter, and D. Wollherr, "A review of spatial reasoning and interaction for real-world robotics," *Adv. Robot.*, vol. 31, pp. 222–242, Jan. 2017.
[6] Y. Tao, L. Ding, and A. Ganz, "Indoor navigation validation framework for visually impaired users," *IEEE Access*, vol. 5, pp. 21763–21773, 2017.
[7] K. Zsiga, A. Tóth, T. Pilissy, O. Péter, Z. Dénes, and G. Fazekas, "Evaluation of a companion robot based on field tests with single older adults in their homes," *Assist. Technol.*, vol. 30, no. 5, pp. 259–266, Oct. 2018.
[8] F. Ali, D. Kwak, P. Khan, S. H. A. Ei-Sappagh, S. M. R. Islam, and D. Park, "Merged ontology and SVM-based information extraction and recommendation system for social robots," *IEEE Access*, vol. 5, pp. 12364–12379, 2017.
[9] J. F. De G. Luengo, F. A. Martín, A. Castro-González, and M. A. Salichs, "Sound synthesis for communicating nonverbal expressive cues," *IEEE Access*, vol. 5, pp. 1941–1957, 2017.
[10] P. Yan, B. He, L. Zhang, and J. Zhang, "Task execution based-on human-robot dialogue and deictic gestures," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Qingdao, China, Dec. 2016, pp. 1918–1923.
[11] Y. Park and S. Kang, "Natural language generation using dependency tree decoding for spoken dialog systems," *IEEE Access*, vol. 7, pp. 7250–7258, 2018.
[12] J. Tan, Z. Ju, and H. Liu, "Grounding spatial relations in natural language by fuzzy representation for human-robot interaction," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Beijing, China, Jul. 2014, pp. 1743–1750.
[13] M. Hanheide, M. Göbelbecker, G. S. Horn, A. Pronobis, K. Sjöö, A. Aydemir, P. Jensfelt, C. Gretton, R. Dearden, M. Janicek, H. Zender, G.-J. Kruijff, N. Hawes, and J. L. Wyatt, "Robot task planning and explanation in open and uncertain worlds," *Artif. Intell.*, vol. 247, pp. 119–150, Jun. 2017.
[14] D. Lu, Y. Zhou, F. Wu, Z. Zhang, and X. Chen, "Integrating answer set programming with semantic dictionaries for robot task planning," in *Proc. IJCAI*, Melbourne, VIC, Australia, Aug. 2017, pp. 4361–4367.
[15] D. K. Misra, J. Sung, K. Lee, and A. Saxena, "Tell me dave: Context-sensitive grounding of natural language to manipulation instructions," *Int. J. Robot. Res.*, vol. 35, nos. 1–3, pp. 281–300, Jan. 2016.
[16] A. Magassouba, K. Sugiura, and H. Kawai, "A multimodal classifier generative adversarial network for carry and place tasks from ambiguous language instructions," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 3113–3120, Oct. 2018.
[17] S. Guadarrama, L. Riano, D. Golland, D. Göhring, Y. Jia, D. Klein, P. Abbeel, and T. Darrell, "Grounding spatial relations for human-robot interaction," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Tokyo, Japan, Nov. 2013, pp. 1640–1647.
[18] T. Kollar, S. Tellex, D. Roy, and N. Roy, "Grounding verbs of motion in natural language commands to robots," in *Proc. Exp. Robot.*, Essaouira, Morocco, 2014, pp. 31–47.
[19] A. S. Huang, S. Tellex, A. Bachrach, T. Kollar, D. Roy, and N. Roy, "Natural language command of an autonomous micro-air vehicle," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Taipei, Taiwan, Oct. 2010, pp. 2663–2669.
[20] R. katsuki, R. Siegwart, J. Ota, and T. Arai, "Reasoning of abstract motion of a target object through task order with natural language—Pre-knowledge of object-handling-task programming for a service robot," *Adv. Robot.*, vol. 20, no. 4, pp. 391–412, Jan. 2006.

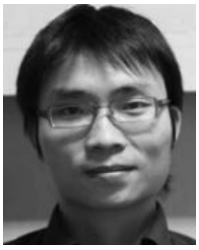
[21] N. Dantam and M. Stilman, "The motion grammar: Analysis of a linguistic method for robot control," *IEEE Trans. Robot.*, vol. 29, no. 3, pp. 704–718, Jun. 2013.
[22] V. Perera and M. Veloso, "Handling complex commands as service robot task requests," in *Proc. IJCAI*, Buenos Aires, Argentina, Jun. 2015, pp. 1177–1183.
[23] J. Thomason, S. Zhang, R. J. Mooney, and P. Stone, "Learning to interpret natural language commands through human-robot dialog," in *Proc. IJCAI*, Buenos Aires, Argentina, Jun. 2015, pp. 1923–1929.
[24] M. A. V. J. Muthugala and A. G. B. P. Jayasekara, "A review of service robots coping with uncertain information in natural language instructions," *IEEE Access*, vol. 6, pp. 12913–12928, 2018.
[25] K. Shuang, Y. Liu, W. Zhang, and Z. Zhang, "Summarization filter: Consider more about the whole query in machine comprehension," *IEEE Access*, vol. 6, pp. 58702–58709, 2018.
[26] K. Yao, B. Peng, G. Zweig, D. Yu, X. Li, and F. Gao, "Recurrent conditional random field for language understanding," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 4077–4081.
[27] X. Zhang and H. Wang, "A joint model of intent determination and slot filling for spoken language understanding," in *Proc. IJCAI*, New York, NY, USA, Jul. 2016, pp. 2993–2999.
[28] B. Liu and I. Lane, "Attention-based recurrent neural network models for joint intent detection and slot filling," in *Proc. Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 685–689.
[29] L. Qiu, Y. Chen, H. Jia, and Z. Zhang, "Query intent recognition based on multi-class features," *IEEE Access*, vol. 6, pp. 52195–52204, 2018.
[30] C.-W. Goo, G. Gao, Y.-K. Hsu, C.-L. Huo, T.-C. Chen, K.-W. Hsu, and Y.-N. Chen, "Slot-gated modeling for joint slot filling and intent prediction," in *Proc. NAACL-HLT*, New Orleans, LA, USA, Jun. 2018, pp. 753–757.
[31] J. Zhou and W. Xu, "End-to-end learning of semantic role labeling using recurrent neural networks," in *Proc. ACL-IJCNLP*, Beijing, China, Jul. 2015, pp. 1127–1137.
[32] J. Nivre, "Dependency parsing," *Lang. Linguistics Compass*, vol. 4, no. 3, pp. 45–86 2006.
[33] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Comput.*, vol. 12, no. 10, pp. 2451–2471, Oct. 2000.
[34] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.
[35] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: https://arxiv.org/abs/1409.0473
[36] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, "The ATIS spoken language systems pilot corpus," in *Proc. Workshop Speech Natural Lang.*, Jun. 1990, pp. 96–101.
[37] P. J. Price, "Evaluation of spoken language systems: The ATIS domain," in *Proc. Workshop Speech Natural Lang.*, Jun. 1990, pp. 91–95.
[38] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The Berkeley framenet project," in *Proc. COLING*, Montreal, QC, Canada, Aug. 1998, pp. 86–90.
[39] C. J. Fillmore, "Frames and the semantics of understanding," *Quaderni di Semantica*, vol. 6, no. 2, pp. 222–254, 1985.
[40] D. Hakkani-Tür, G. Tur, A. Celikyilmaz, Y.-N. V. Chen, J. Gao, L. Deng, and Y.-Y. Wang, "Multi-domain joint semantic frame parsing using bi-directional RNN-LSTM," in *Proc. Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 715–719.

**SHUYOU ZHANG** received the M.S. degree in mechanical engineering and the Ph.D. degree from the State Key Laboratory of CAD&CG, Zhejiang University, China, in 1991 and 1999, respectively.

He is currently a Professor with the Department of Mechanical Engineering, Zhejiang University, China. He is also the Administer of the Institute of Design Engineering, Zhejiang University, an Assistant Director of the Computer Graphics Professional Committee for the China Engineering Graphic Society, a member of the Product Digital Design Professional Committee, and the Chairman of the Zhejiang Engineering Graphic Society. His research interests include human–robot interaction, semantic segmentation, and engineering and computer graphics.

**JUNJIE JIANG** received the B.Sc. degree in process equipment and control engineering from Nanjing Tech University, China, in 2017. He is currently pursuing the Ph.D. degree with the Department of Mechanical Engineering, Zhejiang University, China.

His research interests include spoken language understanding, human–robot interaction, and semantic segmentation.



**ZAIXING HE** received the B.Sc. and M.Sc. degrees in mechanical engineering from Zhejiang University, China, in 2006 and 2008, respectively, and the Ph.D. degree from the Graduate School of Information Science and Technology, Hokkaido University, Japan, in 2012.

He is currently an Associate Professor with the Department of Mechanical Engineering, Zhejiang University. His research interests include deep learning, robotic vision, visual intelligence of manufacturing equipments, and image understanding.

**XINYUE ZHAO** received the M.S. degree in mechanical engineering from Zhejiang University, China in 2008, and the Ph.D. degree from the Graduate School of Information Science and Technology, Hokkaido University, Japan, in 2012.

She is currently an Associate Professor with the Department of Mechanical Engineering, Zhejiang University, China. Her research interests include machine learning and machine vision.

**JINHUI FANG** received the Ph.D. degree in fluid power transmission and control from Zhejiang University, Hangzhou, China, in 2013.

He is currently an Assistant Research Fellow of the State Key Laboratory of Fluid Power and Mechatronic Systems, Zhejiang University, Hangzhou, China. His current research interests include intelligent manufacture and the modeling and control of hydraulic components and electro-hydraulic systems.