Learning Cooperative Persuasive Dialogue Policies using Framing

Takuya Hiraoka[a,*], Graham Neubig[a], Sakriani Sakti[a], Tomoki Toda[b], Satoshi Nakamura[a]

[a] *Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 603-0192, Japan*

[b] *Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan*

Abstract

In this paper, we propose a new framework of cooperative persuasive dialogue, where a dialogue system simultaneously attempts to achieve user satisfaction while persuading the user to take some action that achieves a pre-defined system goal. Within this framework, we describe a method for reinforcement learning of cooperative persuasive dialogue policies by defining a reward function that reflects both the system and user goal, and using framing, the use of emotionally charged statements common in persuasive dialogue between humans. In order to construct the various components necessary for reinforcement learning, we first describe a corpus of persuasive dialogues between human interlocutors, then propose a method to construct user simulators and reward functions specifically tailored to persuasive dialogue based on this corpus. Then, we implement a fully automatic text-based dialogue system for evaluating the learned policies. Using the implemented dialogue system, we evaluate the learned policy and the effect of framing through experiments both with a user simulator and with real users. The experimental evaluation indicates that the proposed method is effective for construction of cooperative persuasive dialogue systems.

Keywords: cooperative persuasive dialogue, framing, reinforcement learning, dialogue modeling, dialogue system

1. Introduction

With the basic technology supporting dialogue systems maturing, there has been more interest in recent years about dialogue systems that move beyond the traditional task-based or chatter bot frameworks. In particular there has been increasing interest in dialogue systems that engage in persuasion or negotiation [1, 2, 3, 4, 5, 6, 7, 8]. In this paper, we propose a method for learning *cooperative* persuasive dialogue systems, in which we place a focus not just on the success of persuasion (the *system* goal) but also user satisfaction (the

*Corresponding author. Tel.: +81 90 5105 5084

Email addresses: takuya-h@is.naist.jp (Takuya Hiraoka), neubig@is.naist.jp (Graham Neubig), ssakti@is.naist.jp (Sakriani Sakti), tomoki@icts.nagoya-u.ac.jp (Tomoki Toda), s-nakamura@is.naist.jp (Satoshi Nakamura)

URL: http://isw3.naist.jp/~takuya-h/ (Takuya Hiraoka)

user goal). This variety of dialogue system has the potential to be useful in situations where the user and system have different, but not mutually exclusive goals. An example of this is a sales situation where the user wants to find a product that matches their taste, and the system wants to successfully sell a product, ideally one with a higher profit margin.

Creating a system that both has persuasive power and is able to ensure that the user is satisfied is not an easy task. In order to tackle this problem with the help of recent advances in statistical dialogue modeling, we build our system upon the framework of reinforcement learning and specifically partially observable Markov decision processes (POMDP) [9, 10], which we describe in detail in Section 2. In the POMDP framework, it is mainly necessary to define a *reward* representing the degree of success of the dialogue, the set of *actions* that the system can use, and a *belief state* to keep track of the system beliefs about its current environment. Once these are defined, reinforcement learning enables the system to learn a policy maximizing the reward.

In this paper, in order to enable the learning of policies for cooperative persuasive dialogue systems, we tailor each of these elements to the task at hand (Section 4):

Reward: We present a method for defining the reward as a combination of the user goal (user satisfaction), the system goal (persuasive success), and naturalness of the dialogue. This is in contrast to research in reinforcement learning for slot-filling dialogue, where the system aims to achieve only the user goal [9, 10], or for persuasion and negotiation dialogues, where the system receives a reward corresponding to only the system goal [1, 2, 3, 4]. We use a human-to-human persuasive dialogue corpus (Section 3, [11]) to train predictive models for achievement of a human persuadee's and a human persuader's goals, and introduce these models to reward calculation to enable the system to learn a policy reflecting knowledge of human persuasion.

System Action: We introduce framing [12], which is known to be important for persuasion, as a system action (i.e., system dialogue act). Framing uses emotionally charged words (positive or negative) to explain particular alternatives. In the context of research that applies reinforcement learning to persuasive (or negotiation) dialogue, this is the first work that considers framing in this way. In this paper the system controls the polarity (positive or negative) and the target alternative of framing (see Table 3 for an example of framing).

Belief State: As the belief state, we use the dialogue features used in calculating the reward function. For example, whether the persuadee has been informed that a particular option matches their preference was shown in human dialogue to be correlated with persuasive success, which is one of the reward factors. Some of the dialogue features reward calculation can not be observed directly by the system, and thus we incorporate them into the belief state.

Based on this framework, we construct the first fully automated text-based cooperative persuasive dialogue system (Section 5). To construct the system, in addition to the policy module, natural language understanding (NLU), and natural language generation (NLG) are required. We construct an NLU module using the human persuasive dialogue corpus and

a statistical classifier. In addition, we construct an NLG module based on example-based dialogue, using a dialogue database created from the human persuasive dialogue corpus.

Using this system, we evaluate the learned policy and the utility of framing (Section 6). To our knowledge, in context of the research for persuasive and negotiation dialogue, it is first time that a learnt policy is evaluated with fully automated dialogue system. The evaluation is done both using a user simulator and real users.

This paper comprehensively integrates our work in [13] and [14], with a more complete explanation and additional experiments. Specifically regarding the additional experimental results, in this paper we additionally perform 1) experimental evaluation using a reward function which exactly matches the learning phase (Section 6.1.1, 6.2), and 2) an evaluation of the effect of NLU error rate (Section 6.1.2).

2. Reinforcement learning

In reinforcement learning, policies are updated based on exploration in order to maximize a reward. In this section, we briefly describe reinforcement learning in the context of dialogue. In dialogue, the policy is a mapping function from a dialogue state to a particular system action. In reinforcement learning, the policy is learned to maximize the reward function, which in traditional task-based dialogue system is user satisfaction or task completion [15]. Reinforcement learning is often applied to models based on the frameworks of Markov decision processes (MDP) or partially observable Markov decision processes (POMDP).

In this paper, we follow a POMDP-based approach. A POMDP is defined as a tuple $\langle S, A, P, R, O, Z, \gamma, b_0 \rangle$ where S is the set of states (representing different contexts) which the system may be in (the system's world), A is the set of actions of the system, $P : S \times A \rightarrow P(S, A)$ is the set of transition probabilities between states after taking an action, $R : S \times A \rightarrow \Re$ is the reward function, O is a set of observations that the system can receive about the world, Z is a set of observation probabilities $Z : S \times A \rightarrow Z(S, A)$, and γ a discount factor weighting longterm rewards. At any given time step i the world is in some unobserved state $s_i \in S$. Because s_i is not known exactly, we keep a hypothesis over states called a belief state b.[1] When the system performs an action $\alpha_i \in A$ based on b, following a policy $\pi : b \rightarrow A$, it receives a reward $r_i(s_i, \alpha_i) \in \Re$ and transitions to state s_{i+1} according to $P(s_{i+1}|s_i, \alpha_i) \in P$. The system then receives an observation o_{i+1} according to $P(o_{i+1}|s_{i+1}, \alpha_i)$. The quality of the policy π followed by the agent is measured by the expected future reward, also called the Q-function, $Q^\pi : b \times A \rightarrow \Re$.

In this framework, we use Neural fitted Q Iteration [16] for learning the system policy. Neural fitted Q Iteration is an offline value-based method, and optimizes the parameters to approximate the Q-function. Neural fitted Q Iteration repeatedly performs 1) sampling

[1]Note that, in this paper we use "belief state" to refer to both 1) known information about a part of the dialogue state (e.g., the most recent system action), and 2) a distribution over all possible hypotheses regarding a part of the dialogue state (e.g., the most recent users' dialogue act). We explain about how we define this belief state in our domain in Section 4.2.3.

training experience using a POMDP through interaction and 2) training a Q-function approximator using training experience. Neural fitted Q Iteration uses a multi-layered perceptron as the Q-function approximator. Thus, even if the Q-function is complex, Neural fitted Q Iteration can approximate the Q-function better than using a linear approximation function. In a preliminary experiment, we confirmed that this is true in our domain as well.
Once the Q-function is learned, the system creates the policy based on the Q-function. In our research, we use the ϵ-greedy policy. Namely, the system randomly selects an action with a probability of ϵ, otherwise selects the action which maximizes the Q-function given the current state.

As Porta et al. noted, (discrete-state) POMDPs can be seen as MDPs with continuous state space that has one dimension per state, which represents the probability of each state in original POMDP [17]. More concretely, assuming the state space of POMDPs is the discrete set $S = \{s_1, ..., s_n, ..., s_N\}$, the state $s_i^{'}$ in corresponding MDPs at time step i can be represented as follows:

$$s_i^{'} = (b_i(s_1), ..., b_i(s_n)...., b_i(s_N)),$$

where b_i represents belief state at turn i. In our paper, we follow that discrete-state POMDPs, and treat it as MDPs with continuous state space. So neural fitted Q iteration should be an appropriate method to solve this problem.

3. Cooperative persuasive dialogue corpus

In this section, we give a brief overview of cooperative persuasive dialogue, and a human dialogue corpus that we use to construct the dialogue models and dialogue system described in later sections. Based on the persuasive dialogue corpus (Section 3.1), we define and quantify the actions of the cooperative persuader (Section 3.2). In addition, we annotate persuasive dialogue acts of the persuader from the point of view of framing (Section 3.3).

3.1. Outline of persuasive dialogue corpus

The cooperative persuasive dialogue corpus [11] consists of dialogues between a salesperson (persuader) and customer (persuadee) as a typical example of persuasive dialogue. The salesperson attempts to convince the customer to purchase a particular product (decision) from a number of alternatives (decision candidates). We define this type of dialogue as "sales dialogue." More concretely, the corpus assumes a situation where the customer is in an appliance store looking for a camera, and the customer must decide which camera to purchase from 5 alternatives. The customer can close the dialogue whenever they want, and choose to buy a camera, not buy a camera, or reserve their decision for a later date.

Prior to recording, the salesperson is given the description of the 5 cameras and instructed to try to convince the customer to purchase a specific camera (the persuasive target). In this corpus, the persuasive target is camera A, and this persuasive target is invariant over all subjects. The customer is also instructed to select one preferred camera from the catalog

Table 1: The beginning of a dialogue from the corpus (translated from Japanese)

Speaker	Transcription	GPF Tag
Cust	Well, I am looking for a camera, do you have camera B? (えーと，カメラがほしいんですけど．) (B カメラってあります？)	PROPQ
Sales	Yes, we have camera B. (あ，B カメラございますよ)	ANSWER
Sales	Did you already take a look at it somewhere? (え，何かでもう調べて来られたんですか？)	PROPQ
Cust	Yes. On the Internet. (あー，そうですね．)	ANSWER
Sales	It is very nice. Don't you think? それいいですよね？)	PROPQ
Cust	Yes, that's right, yes. (はい，そうですね，はい)	INFORM

Table 2: Sytem and user GPF tags

Inform	Answer	Question	PropQ	SetQ	Commisive	Directive

of the cameras[2], and choose one aspect of the camera that is particularly important in making their decision (the determinant). During recording, the customer and the salesperson converse and refer to the information in the camera catalog as support for their utterances.

The corpus includes a role-playing dialogue with participants consisting of 3 salespeople from 30 to 40 years of age and 19 customers from 20 to 40 years of age. All salespeople have experience working in an appliance store. The total number of dialogues is 34, and the total time is about 340 minutes. Table 1 show an example transcript of the beginning of one dialogue. A further example is shown in Table 14 in the appendix.

3.2. Annotation of persuader and persuadee goals

We define the *cooperative persuader* as a persuader who achieves both the persuader and persuadee goals, and *cooperative persuasive dialogue* as a dialogue where both the persuader and persuadee goals have been achieved. To measure the salesperson's success as a cooperative persuader, we annotate each dialogue with scores corresponding to the achievement of the two participants' goals. As the persuader's goal, we use persuasive success measured by whether the persuadee's final decision (purchased camera) is the persuasive target or not. As the persuadee's goal, we use the persuadee's subjective satisfaction as measured by results of a questionnaire filled out by the persuadee at the end of the dialogue: "Evaluate how satisfied you were with the clerk (店員にどれだけ満足したかを 5 段階で評価して下さい)" (1: Not satisfied 3: Neutral 5: Satisfied).

[2]The salesperson is not told this information about customer preferences.

Table 3: An example of positive framing $\langle a_i = \text{A}, p_i = \text{POS}, r_i = \text{NO}\rangle$ (above), and negative framing $\langle a_i = \text{B}, p_i = \text{NEG}, r_i = \text{NO}\rangle$ (below). In these examples, the customer has indicated price as the preferred determinant.

(Camera A is) able to achieve performance of comparable single-lens cameras and can fit in your pocket, this is a point. ((カメラAは) ポケットに入る大きさで一眼並みの性能で撮っていただけるっていうことが，) (今回のポイントなんですけれども)
But, considering the long term usage, you might care about picture quality. You might change your mind if you only buy a small camera (Camera B). (長い間，同じカメラを使っていると、写真の質が気になってくるんです．) (たぶん，小さなカメラ (Camera B) だけ買うと，別のものほしくなってくると思います．)

3.3. Annotated dialogue acts

Each utterance is annotated with two varieties of tags, the first covering dialogue acts in general, and the rest covering framing.

As a tag set to represent traditional dialogue acts, we use general purpose functions (GPF) defined by the ISO international standard for dialogue act annotation [18]. All annotated GPF tags are defined to be one of the tags in this set (Table 2).

More relevant to this work is the *framing* annotation. Framing is the use of emotionally charged words to explain particular alternatives, and is known as an effective way of increasing persuasive power. The corpus contains tags of all instances of negative/positive framing [12, 5], with negative framing using negative words and positive framing using positive words.

The framing tags are defined as a tuple $\langle a, p, r\rangle$ where a represents the target alternative, p takes value NEG if the framing is negative, and POS if the framing is positive, and r is a binary variable indicating whether or not the framing contains a reference to the determinant that the persuadee indicated was most important (for example, the performance or price of a camera). The user's preferred determinant is annotated based on the results of the pre-dialogue questionnaire.

Table 3 shows examples of framing. The example shows positive framing (p=POS) about the performance of Camera A (a=A). In this example, the customer answered that his preference is the price of camera, and this utterance does not contain any description of price. Thus, r=NO is annotated. An example of negative framing about Camera B is also shown below.

The annotation is performed by three human workers:

1. The first worker segments speaker utterances so that one utterance unit is tagged by only one GPF. After that the first worker annotates framing tags for each utterance.
2. The remaining two workers annotate framing and GPF tags without looking at the annotation of each other, and modify segmentation if there are utterances tagged by multiple tags.

For this paper, we re-performed annotation of the framing tags and evaluate inter-annotator agreement, which is slightly improved from Hiraoka et al. [11]. Two annotators

are given the description and examples of tags (e.g. what a positive word is), and practice with these manuscripts prior to annotation. In corpus annotation, at first, each annotator independently chooses the framing sentences. Then, framing tags are independently assigned to all utterances chosen by the two annotators. The inter-annotator agreement of target alternative (a) is 91% (kappa=0.813) , framing polarity (p) is 96.9% (kappa=0.903), reference to alternative (r) is 82% (kappa=0.623)

4. Cooperative persuasive dialogue modeling

In this section, we describe a statistical dialogue model for cooperative persuasive dialogue. The proposed cooperative persuasive dialogue model consists of a user-side dialogue model (Section 4.1) and a system-side model (Section 4.2).

4.1. User simulator

The dialogue model for the user (customer in Section 3) is used to simulate the system's conversational partner in applying reinforcement learning. The user simulator estimates two aspects of the conversation:

1. The user's general dialogue act.
2. Whether the preferred determinant has been conveyed to the user (conveyed preferred determinant; CPD).

The users' general dialogue act is represented using GPF. For example, in Table 1, PROPQ, ANSWER, and INFORM appear as the user's dialogue act. In our research, the user simulator chooses one GPF described in Table 2 or *None* representing no response at each turn. CPD represents that the user has been convinced that the determinant in the persuader's framing satisfies the user's preference. For example, in Table 3, the "performance" is contained in the salesperson's positive framing for camera A. If the persuadee is convinced that the decision candidate satisfies his/her preference based on this framing, we say that CPD has occurred (r=YES)[3]. In our research, the user simulator models CPD for each of the 5 cameras. This information is required to calculate reward described in the following Section 4.2.1. Specifically, GPF and CPD are used for calculating naturalness and persuasive success, which are part of the reward function.

The user simulator is based on an order one Markov chain, and Figure 1 shows its dynamic Bayesian network. The user's GPF G_{user}^{t+1} and CPD C_{alt}^{t+1} at turn $t+1$ are calculated by the following probabilities:

$$P(G_{user}^{t+1}|D^t, U_{eval}) \tag{1}$$

$$P(C_{alt}^{t+1}|C_{alt}^t, F_{sys}^t, G_{sys}^t, U_{eval}). \tag{2}$$

D^t represents a dialogue act of the speaker who is taking a turn at t. If the user is taking a turn, then D^t represents G_{user}^t. In addition, if the system is taking a turn, then D^t

[3]Note that the persuader does not necessarily know if r=YES because the persuader is not certain of the user's preferred determinants.

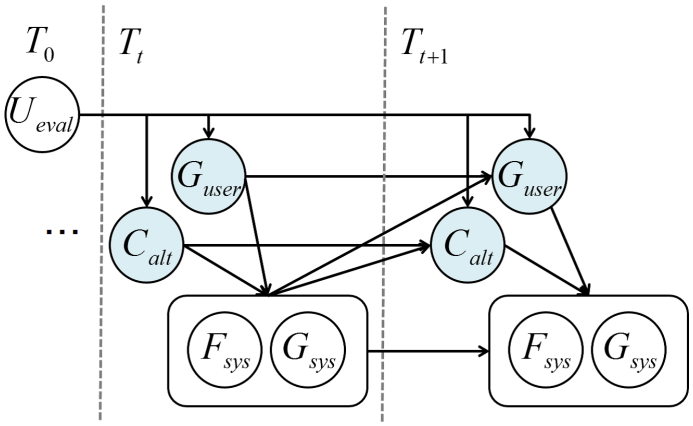


Figure 1: Dynamic Bayesian network of the user simulator. Each node represents a variable, and each edge represents a probabilistic dependency. The system cannot observe the shaded variables.

represents $\{F^t_{sys}, G^t_{sys}\}$. G^t_{sys} represents the system GPF at turn t, and F^t_{sys} represents the system framing at t. These variables correspond to system actions, and are explained in Section 4.2.2. G^t_{user} represents the user's GPF at t, C^t_{alt} represents the CPD at t, and U_{eval} represents the user's original evaluation of the alternatives[4]. In our research, this is the decision candidate that the user selected as a preferred decision candidate at the beginning of the dialogue[5]. Note that a "turn" means one segment of sentences corresponding to one GPF (except for "ReleaseTurn"). More concretely, a row in Table 14 corresponds to one turn. In order to perform mutual turn taking (i.e., the turn belongs to either the user or system), the GPF of the user simulator is ignored in calculation of the simulator's next decision during the system's turn. An example of an application of Equation (1), (2) in simulated dialogue is shown in Table 4 We use the persuasive dialogue corpus described in Section 3 for training the user simulator, considering the customer in the corpus as the user and the salesperson in the corpus as the system.

We use logistic regression for learning Equations (1) and (2). As features, we use a binary vector whose elements correspond to values of G^t_{user}, C^t_{alt}, F^t_{sys}, G^t_{sys}, U_{eval}. We performed an experiment evaluating the quality of the user simulator using leave-one-out cross validation. In this experiment, we evaluate the simulator in terms of its GPF (Equation 1) and CPD (Equation 2) estimation accuracy. Note that, concerning CPD, we evaluate only CPD about camera B because other factors (such CPD about C) do not affect the actual reward calculation.[6] The result (accuracy, precision, recall, F-measure, and perplexity)[7] are

[4]Values of these variables are set at the beginning of dialogue, and invariant over the dialogue.

[5]Preliminary experiments indicated that the user behaved differently depending on the original selection of the decision candidate, thus we introduce this variable to the user simulator.

[6]Note that the system goal is persuading the user to purchase camera A. Our preliminary analysis indicates that informing the user about alternatives (i.e, camera B) other than camera A that match the user's preference increases the system's persuasive power [11].

[7]We use accuracy, precision, recall, and F-measure as evaluation criteria in order to follow Schatzmann

Table 4: An example of an application of Equation (1), (2) in simulated dialogue between the system and the user simulator (only the first 5 turns are shown).

t	Speaker (taking turn)	F	G	Equation (1), (2)
1	System	F^1_{sys} = Pos B	G^1_{sys} = INFORM	$P(G^2_{user}\|F^1_{sys}, G^1_{sys}, U_{eval} = \{CameraB\})$ $P(C^2_{user}\|C^1_{alt} = \{\}, F^1_{sys}, G^1_{sys}, U_{eval} = \{CameraB\})$
2	System	F^2_{sys} = Pos A	G^2_{sys} = INFORM	$P(G^3_{user}\|F^2_{sys}, G^2_{sys}, U_{eval} = \{CameraB\})$ $P(C^3_{user}\|C^2_{alt} = \{CameraB\}, F^2_{sys}, G^2_{sys}, U_{eval} = \{CameraB\})$
3	System	F^3_{sys} = Pos A	G^3_{sys} = INFORM	$P(G^4_{user}\|F^3_{sys}, G^3_{sys}, U_{eval} = \{CameraB\})$ $P(C^4_{user}\|C^3_{alt} = \{CameraB\}, F^3_{sys}, G^3_{sys}, U_{eval} = \{CameraB\})$
	System		RTURN	
4	User (simulator)		G^4_{user} = OTHER	$P(G^5_{user}\|G^4_{user}, U_{eval} = \{CameraB\})$ $P(C^5_{user}\|C^4_{alt} = \{CameraB\}, U_{eval} = \{CameraB\})$
5	System		G^5_{sys} = QUESTION	$P(G^6_{user}\|G^5_{sys}, U_{eval} = \{CameraB\})$ $P(C^6_{user}\|C^5_{alt} = \{CameraB\}, G^5_{sys}, U_{eval} = \{CameraB\})$

Table 5: Quality of the user simulator. The row labeled with "GPF" shows the result of the classification problem for 6 classes (GPFs shown in Table 2). In addition, the row labeled with "CPD (about Camera B)" shows the result of the binary-classification problem. Scores in brackets are those of the baseline. We use simulators that always output majority class in training data as baseline in evaluating accuracy, precision, recall, and F-measure. In addition, we use the simulator that follows the distribution of classes in training data as a baseline in evaluating perplexity. "*" means a significant improvement from the baseline (*: $p < 0.05$, **: $p < 0.01$) according to the t-test.

	Accuracy	Precision	Recall	F-measure	Perplexity
GPF	0.410** (0.370)	0.301 (0.14)	0.410 (0.374)	0.301 (0.204)	4.516** (4.815)
CDP (about Camera B)	0.746 (0.690)	0.739* (0.488)	0.746* (0.698)	0.742* (0.574)	1.773 (1.873)

described in Table 5. These results are similar to learned user simulators in other work [19], we hypothesize that the quality of our simulator is acceptable to use.[8]

4.2. Dialogue modeling: learning cooperative persuasive policies

Now that we have introduced the user model, we describe the system's dialogue management model. In particular, we describe the reward, system action, and belief state, which are required for reinforcement learning.

4.2.1. Reward

We define a reward function according to three factors: user satisfaction, system persuasive success, and naturalness. As the cooperative persuasive dialogue systems must perform

et al. [19]. However, generally there are situations in which there are multiple GPFs that are equally appropriate, and accuracy is an evaluation metric that considers a single GPF only. Therefore, we additionally consider perplexity as an evaluation metric in this evaluation. In perplexity, a distribution of possible GPFs is considered.

[8]The task and experimental conditions (such domain of the dialogue system) in previous work is quite different from those of our work, and thus it is difficult to make a precise comparison.

dialogue to achieve both the system and user goals, we define three elements of the reward function as follows:

Satisfaction (Sat) The user's goal is represented by subjective user satisfaction. The reason why we use satisfaction is that the user's goal is not necessarily clear for the system (and system creator) in persuasive dialogue. For example, some users may want the system to recommend appropriate alternatives, while some users may want the system not to recommend, but only give information upon the user's request. As the goal is different for each user, we use abstract satisfaction as a measure, and leave it to each user how to evaluate achievement of the goal.

Persuasive success (PS) The system goal is represented by persuasive success. Persuasive success represents whether the persuadee finally chooses the persuasive target at the end of the dialogue. Persuasive success takes the value SUCCESS when the customer decides to purchase the persuasive target at the end of dialogue, and FAILURE otherwise.

Naturalness (N) In addition, we use naturalness as one of the rewards. This factor is known to enhance the learned policy performance for real users [20].

We define each of these variables formally as follows. Sat^t_{user} represents a 5 level score of the user's subjective satisfaction (1: Not satisfied, 3: Neutral, 5: Satisfied) at turn t scaled into the range between 0 and 1. PS^t_{sys} represents persuasive success (1: SUCCESS, 0: FAILURE) at turn t. N^t represents bi-gram likelihood of the dialogue between system and user at turn t as follows:

$$N_t = P(F^t_{sys}, G^t_{sys}, G^t_{user} | F^{t-1}_{sys}, G^{t-1}_{sys}, G^{t-1}_{user}). \quad (3)$$

Next, it is necessary to combine these three factors into a single reward function. The importance of each goal will vary depending on the use case of the system. For example, a selfish system could be rewarded with an emphasis on mostly achievement of the system goal, and a cooperative system could be rewarded with an equal emphasis on achievement of both of the goals. However, in the current phase of our research, we have no evidence that one of these factors is more important than the other for cooperative persuasive dialogue, and thus would like to treat them as equally important. Unfortunately, the scale (i.e. the standard deviation) of each factor is different, and thus factors with a larger scale are considered as relatively important, and other factors are considered as relatively unimportant. For example, in our previous research [13], the scale of naturalness N is smaller than other factors, and as a result is largely ignored in the learning. Thus, to ensure that all the factors have an equal influence, we normalize the factors with the z-score.

These 4 normalized factors are then combined into a single reward as follows:

$$r^t = \frac{Sat^t_{user} - \overline{Sat^t_{user}}}{\text{Stddev}(Sat_{user})} + \frac{PS^t_{sys} - \overline{PS_{sys}}}{\text{Stddev}(PS_{sys})} + \frac{N^t - \overline{N}}{\text{Stddev}(N)}, \quad (4)$$

Table 6: Features for calculating reward. These features are also used as the system belief state.

Sat_{user}	Frequency of system commisives
	Frequency of system question
PS_{sys}	Total time
	C_{alt} (for all 6 cameras)
	U_{eval} (for all 6 cameras)
N	System and user current GPF
	System and user previous GPF
	System framing

Table 7: System framing. Pos represents positive framing and Neg represents negative framing. A, B, C, D, E represent camera names.

Pos A	Pos B	Pos C	Pos D	Pos E	None
Neg A	Neg B	Neg C	Neg D	Neg E	

Table 8: System action.

<None, ReleaseTurn>	<None, CloseDialogue>
<Pos A, Inform>	<Pos A, Answer>
<Neg A, Inform>	<Pos B, Inform>
<Pos B, Answer>	<Pos E, Inform>
<None, Inform>	<None, Answer>
<None, Question>	<None, Commissive>
<None, Directive>	

where variables with a bar represent the mean of variables without a bar, and the Stddev function represents standard deviation of the argument.

To evaluate these values automatically, Sat and PS are calculated with a predictive model constructed from the human persuasive dialogue corpus described in Section 3 [11]. In constructing these predictive models, the persuasion results (i.e. persuasive success and persuadee's satisfaction) at the end of dialogue are given as the supervisory signal, and the dialogue features in Table 6 are given as the input. In the reward calculation, the dialogue features used by the predictive model are calculated by information generated from the dialogue of the user simulator and the system. Table 6 shows all features used for reward calculation at each turn[9].

Statistics (i.e. mean and standard deviation of each factor) are calculated from simulated dialogue with the dialogue model proposed in this section. Note that in this simulated dialogue, the system obeys a random policy (i.e. randomly selecting the next system action described in Section 4.2.2). We sampled the reward factor for 60,000 turns of simulated dialogue (about 6000 dialogues) to calculate the statistics of each factor.

4.2.2. Action

The system's action $\langle F_{sys}, G_{sys} \rangle$ is a framing/GPF $\langle a, p \rangle$[10] pair. These pairs represent the dialogue act of the salesperson, and are required for reward calculation (Section 4.2.1). There are 11 types of framing (Table 7), and 9 types of GPF which are expanded by adding ReleaseTurn and CloseDialogue to the original GPF sets (Table 2). The number of all possible GPF/framing pairs is 99, and some pairs have not appeared in the original corpus. Therefore, we reduce the number of actions by filtering. We construct a unigram model of the salesperson's dialogue acts $P(F_{sales}, G_{sales})$ from the original corpus, then exclude

[9]Originally, there are more dialogue features for the predictive model. However as in previous research [11], we choose a subset dialogue features by step-wise feature selection [21].

[10]Note that the r is not included in the framing of system action. We assume the system can not control r because the system is not certain of the user's preferred determinants.

pairs for which the likelihood is below 0.005[11]. As a result, the 13 pairs shown in Table 8 remained[12]. We use these pairs as the system actions.

4.2.3. Belief state

The current system belief state is represented by the features used for reward calculation (Table 6) and the reward calculated at the previous turn. Namely, the features for the reward calculation and calculated reward are also used as the next input of the system policy.

Note that the system cannot directly observe C_{alt}, thus the system estimates it through the dialogue by using the following probability.

$$P(\hat{C_{alt}^{t+1}}|F_{sys}^{t}, G_{sys}^{t}, U_{eval}) = \sum_{\hat{C_{alt}^{t}}} P(\hat{C_{alt}^{t+1}}|\hat{C_{alt}^{t}}, F_{sys}^{t}, G_{sys}^{t}, U_{eval}) P(\hat{C_{alt}^{t}}). \quad (5)$$

where $\hat{C_{alt}^{t+1}}$ represents the estimated CPD at $t+1$, and $\hat{C_{alt}^{t}}$ represents the estimated CPD at t. The other variables are the same as those in Equation (2).

In addition, the system also cannot directly observe G_{user}, thus the system estimates it through the dialogue by using the following equation.

$$P(G_{user}^{t+1}|H_{G_{user}^{t+1}}) = \frac{\sum_{G_{user}^{t}} P(H_{G_{user}^{t+1}}|G_{user}^{t+1}) P(G_{user}^{t+1}|G_{user}^{t}) P(G_{user}^{t})}{\sum_{G_{user}^{t+1}} \sum_{G_{user}^{t}} P(H_{G_{user}^{t+1}}|G_{user}^{t+1}) P(G_{user}^{t+1}|G_{user}^{t}) P(G_{user}^{t})}. \quad (6)$$

$H_{G_{user}^{t+1}}$ represents the NLU result (described in Section 5.1) at t. Other variables are the same as those in Eq. (1) and Eq. (2). $P(H_{G_{user}^{t+1}}|G_{user}^{t+1})$ represents the confusion matrix. To construct the confusion matrix, in Section 5.1, we perform an evaluation of NLU and use the confusion matrix from this evaluation for the estimation of Eq. (6). $P(G_{user}^{t+1}|G_{user}^{t})$ is constructed with the persuasive dialogue corpus described in Section 3.1.

The system uses these estimated distributions over the above information (i.e., C_{alt} and G_{user}) in order to determine its next action. Note that other features, which are also described in Table 6, are not estimated as a distribution.

5. Text-based cooperative persuasive dialogue system

To evaluate the policy learned with the dialogue model described in Section 4, we construct a fully automated text-based cooperative persuasive dialogue system. The structure of the system is shown in Figure 2. Especially, in this section, we describe the construction of NLU (Section 5.1) and NLG (Section 5.2) modules that act as an interface between the policy module and the human user, and are necessary for fully automatic dialogue.

[11]We chose this threshold by trying values from 0.001 to 0.01 with incrementation of 0.001. We select the threshold that resulted in the number of actions closest to previous work [2].

[12]Cameras C and D are not popular, and don't appear frequently in the human persuasive dialogue corpus, and are therefore excluded in filtering.

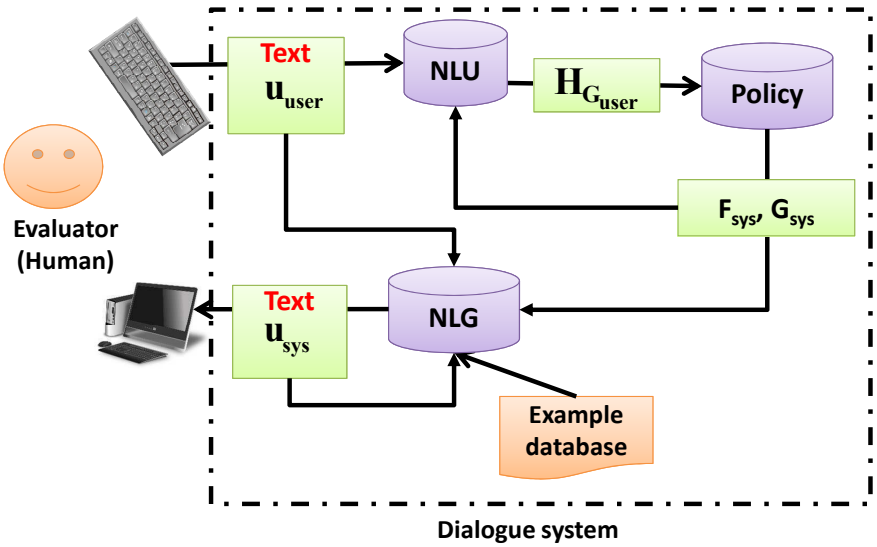


Figure 2: Structure of our dialogue system. Rectangles represent information, and cylinders represent a system module.

5.1. Natural language understanding

The NLU module detects the GPF in the user's text input u_{user} using a statistical classifier. In this paper, we use bagging, using decision trees as the weak classifier [22]. We require the NLU to 1) be simple and 2) output the estimated classes with probability, and bagging with decision trees satisfies these requirements. The NLU uses many features (i.e. word frequency), and decision trees can select a small number of effective features, making a simple classifier. In addition, by using bagging, the confidence probability, which is determined by the voting rate of decision trees, can be attached to the classification result. We utilize Weka [23] for constructing the bagging classifier.

As input to the classifier, we use features calculated from u_{user} and the history of system outputs (u_{sys}, $\langle G_{sys}, F_{sys} \rangle$). Features are mainly categorized into 4 types:

Uni: Unigram word frequency in the user's input.

Bi: Bigram word frequency in the user's input.

DAcl: The previous action of the system (i.e. GPF/framing pairs $\langle G_{sys}, F_{sys} \rangle$).

Unicl: Unigram word frequency in the previous system utterance.

As we use Japanese as our target language, we perform morphological analysis using Mecab [24], and use information about the normal form of the word and part of speech to identify the word.

As the NLU result $H_{G_{user}}$, 8 types of GPF are output with membership probabilities. We use 694 customer utterances in the camera sales corpus (Section 3) as training data. In this training data, 8 types of GPF labels are distributed as shown in Table 9.

Table 9: Distribution of the GPF labels in the training data.

Other	Question	SetQuestion	PropositionalQuestion	Inform	Answer	Directive	Commissive
46	4	12	156	260	117	36	63

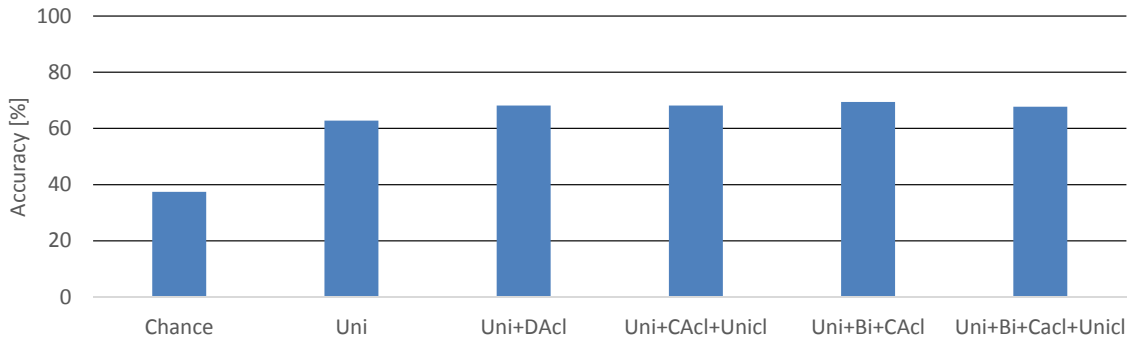


Figure 3: Accuracy of the NLU module. The vertical axis represents accuracy and the horizontal axis represents the NLU feature set. Chance rate is an NLU module that always outputs Inform.

Table 10: The confusion matrix. Each row represents the distribution of the true GPF label. Each column represents the distribution of the NLU classification result.

Other	Commissive	PropQ	Directive	Answer	Inform	SetQ	Question	Classified as/True label
43	0	0	0	0	3	0	0	Other
6	31	2	4	0	20	0	0	Commssive
0	1	112	3	0	40	0	0	PropQ
2	2	6	13	0	13	0	0	Directive
0	3	5	0	53	56	0	0	Answer
1	12	4	4	9	230	0	0	Inform
0	0	10	0	0	2	0	0	SetQ
0	0	3	0	0	1	0	0	Question

We evaluate the performance of the NLU module using the features shown above. We prepare 4 patterns of feature sets (Uni, Uni+DAcl, Uni+CAcl+Unicl and Uni+CAcl+Bi), and evaluate the NLU module with respect to recognition accuracy of GPF labels in the customer's utterances. The evaluation is performed based on 15-fold cross-validation with 694 customer utterances.

From the experimental result (Figure 3), we can see that NLU with Uni+CAcl+Bi achieves the highest accuracy, and thus we decided to use Uni+CAcl+Bi for NLU of the dialogue system in Section 6. Focusing on the details of the misclassified GPFs, we show the confusion matrix for classification results of the NLU module with Uni+CAcl+Bi in Table 10. From this matrix, we can see that Answer is misclassified to Inform, and that SetQ and Question are misclassified into PropositionalQ. This result indicates that this module has difficulty in distinguishing dialogue acts in a hypernym/hyponym or sibling relationship.

5.2. Natural language generation

Our natural language generation module produces a system utterance utilizing surface information of the previous system utterance or user utterance. Note that our dialogue

Table 11: Part of the example database. The sentences surrounded by <> are inserted in correction.

Speaker	Utterance	GPF	Framing
Sys.	What was the good point of camera A? (Aのカメラのどこがよかったんですか?)	Question	
User	Well, I like its shape, like a Monolith. (そうですね。このモノリスみたいな露骨な形が好だからです)	Answer	
Sys.	The <main> difference between camera A <and other cameras> is the sensor. (Aのカメラ < と他のカメラの大きな > 違いはセンサーです) It is said that sensors are essential for a digital camera. (デジタルカメラはセンサーが命といわれています) The sensor of camera A is the same as that as a single-lens cameras. (Aのカメラのセンサーは一眼と同じセンサーを使ってるんですね。)	Inform	Pos A
Sys.	In addition, the size of A is similar to other cameras. (なお、Aのカメラの大きさは他のカメラと一緒です。)	Inform	
User	That's great. (それはすごいですね)	Inform	

model (Section 4) and natural language understanding module (Section 5.1) consider the illocutionary force aspect of utterances (such as "Inform", "Answer", and "Question" in GPF), but do not consider semantic content (such as topics of "Question" in GPF, and attibutes of camera in framing) explicitly. Instead, in the natural language generation module (Section 5.2), the system utterances are generated considering *approximated'* semantic content (of both user and system utterance) in order to achieve semantically coherent dialogue. Specifically, it utilizes n-gram and other surface features of speakers utterances in order to approximate semantic content.

The NLG module outputs a system response u_{sys} based on the user's input u_{user}, the system's previous utterance u'_{sys} and the system action $\langle G_{sys}, F_{sys} \rangle$. Though the dialogue assumed in this paper is focusing on a restricted situation, it is still not trivial to create system responses for various inputs. In order to avoid the large amount of engineering required for template-based NLG and allow for rapid prototyping, we decide to use the framework of example-based dialogue management [25].

We construct an example database $D = \{d_1, d_2, ..., d_M\}$ with M utterances by modifying the human persuasive dialogue corpus of Section 3. In the example database, the ith datum $d_i = \langle s, u, g, f, p \rangle$ consists of the speaker s, utterance u, GPF g, framing flag f, and previous datum p. In modifying the human persuasive dialogue corpus, we manually make the following corrections:

- Deletion of redundant words and sentences (e.g. fillers and restatements).
- Insertion of omitted words (e.g. subjects or objects) and sentences.

Our example database consists of 2022 utterances (695 system utterances and 1327 user example utterances). An example of the database is shown in Table 11.

The NLG module determines the system response u_{sys} from D, considering u_{user}, u'_{sys}, and $\langle G_{sys}, F_{sys}\rangle$. More concretely, our NLG module performs the following procedure:

1. We define the response candidate set R, which is a subset of D, according to whether u_{user} is null $\varnothing$ or not. If $u_{user} \neq \varnothing$ (i.e., the user spoke to the system most recently), then we define R as the set of utterances r for which the previous utterance is a user utterance ($r.p.s = User$) and annotated with the GPF estimated by NLU ($r.p.g = \arg\max_{G_{user}} H_{G_{user}}$). Conversely, if $u_{user} = \varnothing$ (i.e., the system spoke to the user most recently, and is continuing speaking), then we define R so $r.p.s = Sys$,$r.p.g = G'_{sys}$, and $r.p.f = F'_{sys}$, where G'_{sys} represents GPF and F'_{sys} represents framing in the previous system action.[13]
2. Response candidates R are scored based on the following similarity score

$$\cos(r.p.u, u_{input}) = \frac{\text{words}(r.p.u) \cdot \text{words}(u_{input})}{\mid \text{words}(r.p.u) \mid \cdot \mid \text{words}(u_{input}) \mid} \tag{7}$$

$$u_{input} = \begin{cases} u'_{sys} & (u_{user} = \varnothing) \\ u_{user} & (u_{user} \neq \varnothing). \end{cases}$$

   The cosine similarity cos between the previous utterance of the response sentence candidate $r.p.u$ ($r \in R$) and input sentence u_{input} is used for the scoring. u_{input} is set as u'_{sys} or u_{user} depending on u_{user}. The words function returns the frequency vector of the content words (i.e. nouns, verbs, and adjectives) weighted according to tf-idf.
3. The $r^*.u$ that has the highest score is selected as the output of the NLG module u_{sys}

$$r^* = \arg\max_{r \in R} \cos(r.p.u, u_{input}) \tag{8}$$

$$u_{sys} = r^*.u. \tag{9}$$

Note that the language generation is used for generating the actual system utterance corresponding to system action $\langle F_{sys}, G_{sys}\rangle$, and that the decision whether a system should speak more or wait is determined by the system policy described in Section 4.2. If the system selects "<None, ReleaseTurn>" in Table 8, the corresponding system utterance "How about it? (いかがでしょうか？)" is generated as the system utterance, and then the system waits for the user response. Otherwise the system keeps speaking.

6. Experimental evaluation

In this section, we describe the evaluation of the proposed method for learning cooperative persuasive dialogue policies. Especially, we focus on examining how the learned policy with framing is effective for persuasive dialogue. The evaluation is done both using a user simulator (Section 6.1) and real users (Section 6.2).

[13] In this paper, we use "." for representing the membership relation between variables. For example, $Var1.Var2$ means that $Var2$ is a member variable of $Var1$.

6.1. Policy learning and evaluation using the user simulator

In this section, we perform two types of evaluation. At first, we evaluate the effectiveness of framing and learning policies with the user simulator (Section 6.1.1). We also perform an evaluation of how NLU performance affects the learning of the dialogue policy (Section 6.1.2).

6.1.1. Evaluation for the learned policy and framing

For evaluating the effectiveness of framing and learning the policy through the user simulator, we prepare the following 3 policies.

Random: A baseline where the action is randomly output from all possible actions.

NoFraming: A baseline where the action is output based on the policy which is learned using only GPFs. For constructing the actions, we remove actions whose framing is not *None* from the actions described in Section 4.2.2. The policy is a greedy policy, and selects the action with the highest Q-value.

Framing: The proposed method where the action is output based on the policy learned with all actions described in Section 4.2.2 including framing. The policy is also a greedy policy.

For learning the policy, we use Neural fitted Q Iteration (Section 2) using the Pybrain library [26]. We set the discount factor γ to 0.9, and the number of nodes in the hidden layer of the neural network for approximating the Q-function to the sum of number of belief states and actions (i.e. Framing: 53, NoFraming: 47). The policy in learning is the ϵ-greedy policy ($\epsilon = 0.3$). These conditions follow the default Pybrain settings. We consider 2000 dialogues as one epoch, and update the parameters of the neural network at each epoch. Learning is finished when number of epochs reaches 20 (40000 dialogues), and the policy with the highest average reward is used for evaluation.

We evaluate the system on the basis of average reward per dialogue with the user simulator. For calculating average reward, 1000 dialogues are performed with each policy[14].

Experimental results (Figure 4) indicate that 1) performance is greatly improved by learning and 2) framing is somewhat effective for the user simulator. Learned policies (Framing, NoFraming) get a higher reward than Random. Particularly, both of the learned policies achieve better user satisfaction and naturalness than Random. In addition, reward of Framing is higher than NoFraming, specifically because framing is effective for persuasive success. On the other hand, user satisfaction of Framing is lower than that of NoFraming, indicating that there is some tradeoff between user satisfaction and other factors.

[14]We also optimized the policy in the case where the reward (Equation (4)) is given only when dialogue is closed. However, the learning did not converge well, and thus we use the reward (Equation (4)) instead. The use of rewards that are given incrementally, like Equation (4), to improve learning speed and convergence is called "Reward shaping" in reinforcement learning literature (see [27], [28] for the detail). The convergence of the learning in each reward condition is shown in Figures 9 in the appendix.

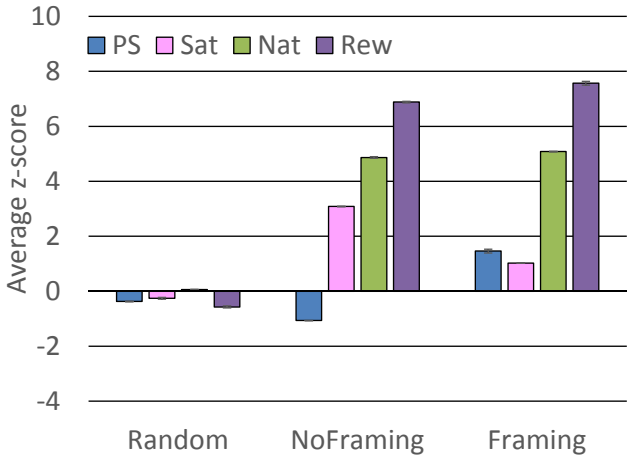


Figure 4: Average value of reward (on z-scaled scale described in Section 4.2.1) for dialogue with the simulator. Error bars represents 95% confidence intervals. Rew represents the reward, Sat represents the user satisfaction, PS represents persuasive success, and Nat represents naturalness.

6.1.2. Evaluation of the effect of NLU performance

To elucidate how the performance of NLU affects the learning of the policy, we prepare 4 Framing policies with different NLU error rates (**Err0%**, **Err25%**, **Err50%**, **Err75%**). These policies are basically the same as the Framing policy in the previous section. However, outputs of the NLU module in these policies contains errors based on a confusion matrix with an overall error rate corresponding to their name (e.g the NLU error rate in Err25% is 25%). These confusion matrices are randomly created at the beginning of each dialogue. We use Neural fitted Q Iteration, whose learning parameters (i.e γ and ϵ), and number of epochs and dialogues are the same as the previous section.

We evaluate the system on the basis of average reward per dialogue with the user simulator. For calculating average reward, 1000 dialogues are performed with 20 learned policies at each error level. In addition, we investigate the informativeness of the estimated GPF distribution by calculating entropy (i.e Eq. (6)).

Experimental results (Figure 5, 6) indicate that average rewards reach the minimum value with the policy where the estimated GPF reaches the highest average entropy. Focusing on the average reward of each system (Figure 5), the average reward of Err75% is smallest of policies, and the average reward gradually decreases as the error rate of policies approaches 75%. In addition, focusing on the average entropy (Figure 6), the average entropy of the estimated GPF reaches the highest value at Err75%, and its value gradually decreases as the system error rates decrease from Err75%. These results indicate that there is a correlation between the performance of NLU and overall evaluation (reward) of the system in our persuasive dialogue model. Note that, in experimental evaluations in other sections, we use NLU constructed in the Section 5.1. This NLU error rate is about 30%, and we can expect that the average reward of this system will be close to that of Err25%.

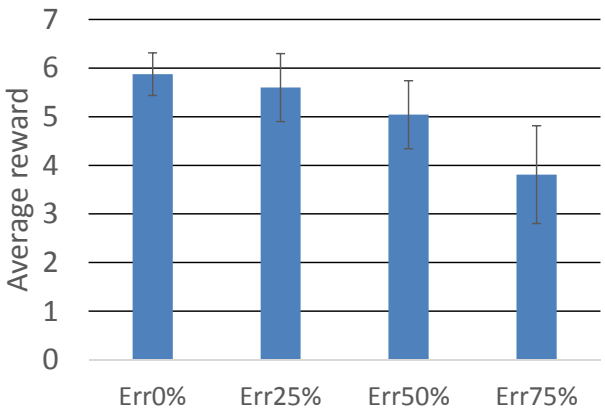


Figure 5: Average reward for dialogue with the user simulator. Error bars represents 95% confidence intervals.

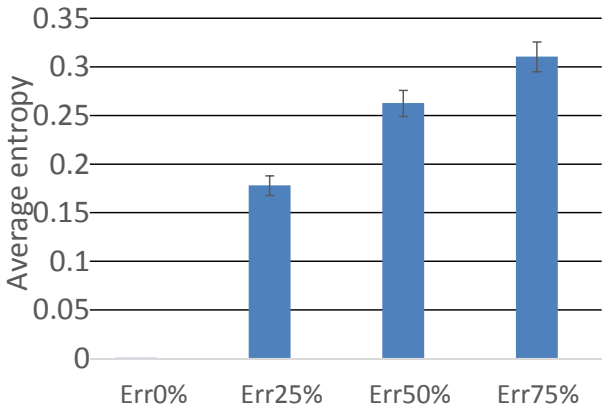


Figure 6: Average entropy of the GPF distribution. Error bars represents 95% confidence intervals.

6.2. Complete system evaluation with real users

To test whether the gains shown on the user simulator will carry over to an actual dialogue system, we perform an experiment with real human users. In this section, we describe the results of our user study evaluating fully automated cooperative persuasive dialogue systems. The system follows the structure proposed in Section 5. The purpose of this is experimental comparison of policies learnt over simulated dialogue (Section 6.1.1) and an actual human policy. We hypothesize that Framing in Section 6.1.1 is the best policy among the learnt policies, and comparable to Human in terms of rewards defined by Equation (4).

For evaluation, in addition to the policies described in Section 6.1.1, we add the following policy.

Human An oracle where the action is output based on human selection. In this research, the first author (who has no formal sales experience, but experience of about 1 year in analysis of camera sales dialogue) selects the action.

We evaluate policies on the basis of average reward and correct response rate of dialogues with real users. The definition of the reward is described in Section 4.2.1. In addition, the correct response rate is the ratio of correct system responses to all system responses. In the experiment, the dialogue system proposed in Section 5 plays the salesperson, and the user plays the customer. At the end of the dialogue, to calculate the reward, the user answers the following questionnaire:

Satisfaction: The user's subjective satisfaction defined as a 5 level score of customer satisfaction (1: Not satisfied，3: Neutral，5: Satisfied).

Final decision: The camera that the user finally wants to buy.

In addition, to calculate the correct response rate, we have the user annotate information regarding whether each system response is correct or not. When we instructed each user about the annotation, we simply ask them to "Mark system responses that seem incorrect to you (あなたにとって，正しくないシステムの発話に印をつけてください．)". An example

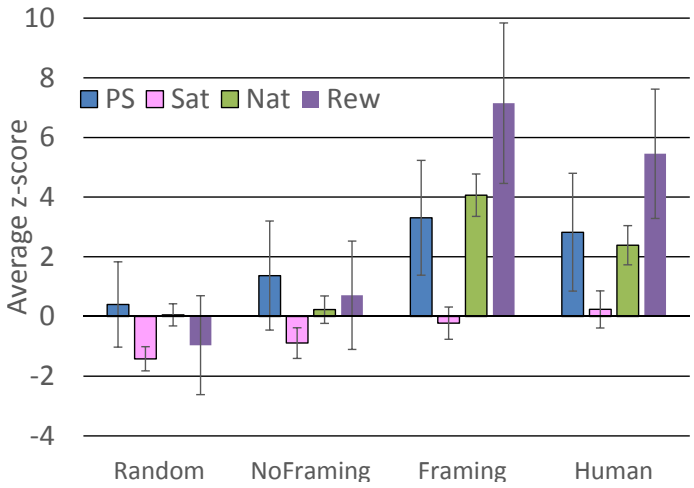


Figure 7: Average value of reward (on z-scaled scale described in Section 4.2.1) for dialogue with real users. Error bars represent 95% confidence intervals. Rew represents the reward, Sat represents the user satisfaction, PS represents persuasive success, and Nat represents naturalness.

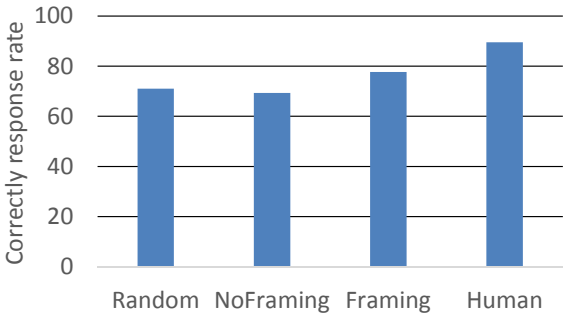


Figure 8: Correct response rate of the system utterances.

of correct/incorrect system responses is shown in Table 12. 13 users perform one dialogue with the system obeying each policy (a total of 4 dialogues per users).

Experimental results for the reward are shown in Figure 7. These results indicate that learning a policy with framing is effective in the text-based cooperative dialogue system[15]. We can see that the reward of Framing is higher than that of NoFraming and Random, and not statistically different from Human. The naturalness of Human is less than that of Framing. One of the reasons for this is that naturalness is automatically evaluated by Equation (3), and Framing is learnt considering this objective naturalness because it is included in the reward (i.e., equation (4)). In contrast to Framing, Human does not carefully focus on increasing this objective naturalness given by Equation (3), resulting in it scoring lower objective naturalness than Framing. Experimental results for the correct response rate (Figure 8) indicate that our cooperative persuasive dialogue system somewhat correctly responds to the user's input. The scores of all policies are higher than 70%, and the score of Framing is about 77%. In addition, even the Random policy achieves a score of about 70%. One of the reasons for this is that NLG method used by our system (Section 5.2) is based on examples, and thus is able to return natural responses that will only be judged as incorrect if they do not match the context.

We can see that some features in human persuasive dialogue appear in the dialogue between users and the system obeying the Framing policy. An example of a typical dialogue of Framing is shown in Table 12 (original Japanese transcription is shown in Table 13). The first feature is that most of the framing that the system performs in the dialogue is positive framing for camera A. Even when the user asks about other topics (e.g. camera B∼E and

[15]Note that scores in Figure 7 are normalized into z-score (see Equation (4) in Section 4.2.1). The mean of user satisfaction in the original scale is 3.18 (close to "Neutral"), and thus the value zero in user satisfaction in Figure 7 is equal to 3.18 in the original scale.

determinants) which are not camera A, the system tries to perform framing for camera A. This feature commonly appeared in the human persuasive dialogue. The second feature is that the system checks or asks about user's profile and their thoughts before performing framing. This feature is often found in human dialogue when the user satisfaction is high. In contrast to these features, there are some feature which do not appear in human dialogue. One of the features is that the system talks much more than the user. In the dialogue, most of the dialogue is occupied with system dialogue, and the number of user utterances is very small (on average of about 3 or 4 utterance). One of the reasons for this is that the reward for the system is determined according to estimated rewards on a human corpus, which use the features in Table 6. It can be noted from this table that there is no feature other than total time preventing the system from being overly verbose, largely due to the fact that none of the human persuaders used in the training data showed this kind of behavior. This indicates that we might potentially get further improvements in the system by using data not only from human-human interactions, but also from human-computer interactions in the calculation of the reward function.

Considering the evaluation result of Section 6.1.1, we can see that trend of reward and its factors differs somewhat between the user simulator and the real users. While the naturalness and reward of Framing are identical in Figures 4 and 7, the systems are given excessively high Sat in simulation. In addition, systems are given underestimated PS in simulation. One of the reasons for this is that the property of dialogue features for the predictive model for reward differs from previous research [11]. In this paper, dialogue features for the predictive model are calculated at each turn. In addition, persuasive success and user satisfaction are successively calculated at each turn. In contrast, in previous research, the predictive model was constructed with dialogue features calculated at end of the dialogue. Therefore, it is not guaranteed that the predictive model estimates appropriate persuasive success and user satisfaction at each turn[16]. Another reason is that the simulator is not sufficiently accurate to use for reflecting real user's behavior. Compared to other works [29, 30], we are using a relatively small sized corpus for training the user simulator. Therefore, improving the user simulator is an important challenge for future work.

In this Section (and the previous section), we performed evaluation of each policy, where we compare NoFraming and Framing to elucidate the effect of framing. In the current set of these evaluations, Framing has two advantages over NoFraming: first, it has an awareness of the framing feature of actions, and second, it has more actions to choose from. It is possible that some of the gain observed is from the fact that the framing actions are better, independent of the framing feature. Detailed evaluation of the effect of use of framing is a potential avenue for future work.

[16]However, the learning of the policy converges better than in the condition where the reward is given only at the end of dialogue (Figures 9). Therefore, we believe giving reward at each turn is an effective way to ensure the policy is learned robustly.

Table 12: Part of a dialogue between Framing and an evaluator (translated from Japanese). Note that user's GPF is estimated. In this example, the user prefers camera E at the beginning of the dialogue. The system utterance with bold font was regarded as an incorrect response by the user.

Speaker	Transcription	Fra	GPF
System	Shutter-speed of camera A is faster than that of camera B. Have you ever taken a picture and had it turn out different from what you want? For example, when you take a picture of a face, and the face in the picture looks quite different. If the shutter-speed is high, the camera can capture the object in a moment.	Pos A	INFORM
System	The main difference between camera A and other cameras is the sensor. It is said that sensors are essential for a digital camera. The sensor of camera A is the same as that of a single-lens cameras.	Pos A	INFORM
System	In addition, the size of A is similar to other cameras.	None	INFORM
System	(Camera A is) able to achieve performance of comparable single-lens cameras and can fit in your pocket, this is a point.	Pos A	INFORM
System	Camera A is most popular in Kakaku.com now.	Pos A	INFORM
System	Camera E is lightweight.	Pos E	INFORM
System	But, camera A is very popular now.	Pos A	INFORM
System	The resolution of (Camera) A is 2090 million pixels, and zoom is 3.6 times. (Camera) A uses a very big lens, and processes the picture faster and with high resolution.	Pos A	INFORM
System	How about it?	None	RTURN
User	Is (camera) A more heavy than (camera) E?	None	PROPQ
System	**Which pictures do you want to take? Far or near?**	None	PROPQ
System	Considering the long term usage, you might care about picture quality. Camera A will satisfy such your need. There are basically no cameras similar in performance to camera A.	Pos A	INFORM
System	I hope that, at first, you look up all of the cameras, and then decide which camera you want. But, Camera A is definitely most popular.	Pos A	INFORM
System	How about it?	None	RTURN
User	Which one is more reasonably priced? A or B?	None	PROPQ
System	If you buy the cheap camera at first and re-buy new one, the total price is expensive. Right?	None	QUESTION
System	The general price of a single-lens is about 100 or 200 thousand yen. Considering these prices, camera A is a good choice.	Pos A	INFORM
System	How about it?	None	RTURN
User	(Camera) A seems to be good.	None	INFORM
	⟨Dialogue is closed by the user ⟩		

7. Related work

There are a number of related works that apply reinforcement learning to persuasion and negotiation dialogue. Georgila and Traum [1] apply reinforcement learning to negotiation dialogue using user simulators divided into three types representing individualist, collectivist, and altruist. Dialogue between a florist and a grocer is used as an example of negotiation dialogue. In addition, Georgila [2] also applies reinforcement learning to two-issue negotiation dialogue where participants have a party, and decide both the date and food type. A handcrafted user simulator is used for learning the policy of each participant. Heeman [4] models negotiation dialogue, assuming a furniture layout task, and Paruchuri et al. [3] model negotiation dialogue, assuming the dialogue between a seller and buyer.

Our research differs from these in three major ways. The first is that we use framing, positive or negative statements about the particular item, which is known to be important for persuasion [12]. By considering framing, the system has the potential to be more persuasive.

While there is one previous example of persuasive dialogue using framing [6], this system does not use an automatically learned policy, relying on handcrafted rules. In contrast, in our research, we apply reinforcement learning to learn the system policy automatically.

In addition, in these previous works, rewards and belief states are defined with heuristics. In contrast, in our research, reward is defined on the basis of knowledge of human persuasive dialogue. In particular, we calculate the reward and belief state using the predictive model of Hiraoka et al. [11] for estimating persuasive success and user satisfaction using dialogue features. In the real world, it is unclear what factors are important for achieving the dialogue goal in many persuasive situations. By considering these predictions as knowledge of human persuasion, the system can identify the important factors in human persuasion and can track the achievement of the goal based on these.

Finally, these works do not evaluate the learned policy, or evaluate only in simulation. In contrast, we evaluate the learned policy with real users.

8. Conclusion

In this paper, we applied reinforcement learning for learning cooperative persuasive dialogue system policies using framing, and evaluated the learned policies with a fully automated dialogue system. In order to apply reinforcement learning, a user simulator and reward function were constructed based on a human persuasive dialogue corpus. Then, we implemented a fully automatic dialogue system for evaluating the learned policies. We evaluated the learned policy and effect of framing using the constructed dialogue system, a user simulator and real users. Experimental evaluation indicates that applying reinforcement learning is effective for construction of cooperative persuasive dialogue systems that use framing.

In the future, we plan to evaluate the system policies in more realistic situations, that move beyond role-playing to real sales situations over more broad domains. In this research, corpus collection and evaluation are performed in a role-playing situation. Therefore, we plan to evaluate the system policies in a real sales scenario such as in a store with actual customers.

Further, we plan to collect additional corpora in several domains and conditions allowing us to broaden the domains to which the proposed method can be applied. Perhaps the most important avenue of future work is improving our current dialogue model (especially, framing in the system action) to be more general and portable. In our current dialogue model, the system can not persuade the user to make decisions (e.g, purchase a camera) that did not appear in the training data. One way for dealing with this problem is considering features of decision candidates (e.g., price of camera) instead of the decision candidate itself in system framing. To do so, we plan to consider an argumentation framework that uses features of decision candidate [31, 32] into the current framing framework.

We also plan to consider multimodal information [33] for estimating persuasive success and user satisfaction. We plan on collecting a multimodal corpus that includes such non-verbal information, and expand our dialogue model to consider this information.

In addition, there is an open problem about the best method to measure for achievement of the user's goal. In this paper, we use users' (persuadees') subjective satisfaction (i.e., how much the users were satisfied with the systems (persuaders)) to quantify achievement of user goal (following previous work [15]), but there are many other alternatives. One alternative is directly evaluating user goal satisfaction (i.e., whether the user could purchase a camera he/she likes or not). This "user goal satisfaction" has the potential to measure achievement of the user's goal more accurately in our task. However determining which measurement is the best for practical use is difficult, and still an open problem.

Acknowledge

Part of this research was supported by JSPS KAKENHI Grant Number 24240032 and the Commissioned Research of National Institute of Information and Communications Technology (NICT), Japan.

References

1. Georgila, K., Traum, D.. Reinforcement learning of argumentation dialogue policies in negotiation. *Proceedings of INTERSPEECH* 2011;.
2. Georgila, K.. Reinforcement learning of two-issue negotiation dialogue policies. *Proceedings of the SIGDIAL* 2013;.
3. Paruchuri, P., Chakraborty, N., Zivan, R., Sycara, K., Dudik, M., Gordon, G.. POMDP based negotiation modeling. *Proceedings of the first MICON* 2009;.
4. Heeman, P.A.. Representing the reinforcement learning state in a negotiation dialogue. *Proceedings of ASRU* 2009;.
5. Mazzotta, I., de Rosis, F.. Artifices for persuading to improve eating habits. *AAAI Spring Symposium: Argumentation for Consumers of Healthcare* 2006;.
6. Mazzotta, I., de Rosis, F., Carofiglio, V.. PORTIA: a user-adapted persuasion system in the healthy-eating domain. *Intelligent Systems* 2007;.
7. Nguyen, H., Masthoff, J., Edwards, P.. Persuasive effects of embodied conversational agent teams. *Proceedings of HCI* 2007;.
8. Guerini, M., Stock, O., Zancanaro, M.. Persuasion model for intelligent interfaces. *Proceedings of the IJCAI Workshop on Computational Models of Natural Argument* 2003;.
9. Levin, E., Pieraccini, R., Eckert, W.. A stochastic model of human-machine interaction for learning dialog strategies. *Proceedings of ICASSP* 2000;.
10. Williams, J.D., Young, S.. Scaling POMDPs for spoken dialog management. *IEEE Transactions on Audio, Speech, and Language Processing* 2007;.
11. Hiraoka, T., Neubig, G., Sakti, S., Toda, T., Nakamura, S.. Construction and analysis of a persuasive dialogue corpus. *Proceedings of IWSDS* 2014;.
12. Irwin, L., Schneider, S.L., Gaeth, G.J.. All frames are not created equal: A typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes 762* 2013;.
13. Hiraoka, T., Neubig, G., Sakti, S., Toda, T., Nakamura, S.. Reinforcement learning of cooperative persuasive dialogue policies using framing. *Proceedings COLING* 2014;.
14. Hiraoka, T., Neubig, G., Sakti, S., Toda, T., Nakamura, S.. Evaluation of a fully automatic cooperative persuasive dialogue system. *Proceedings of IWSDS* 2015;.
15. Walker, M.A., Litman, D.J., Kamm, C.A., Abella, A.. PARADISE: a framework for evaluating spoken dialogue agents. *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics* 1997;.

16. Riedmiller, M.. Neural fitted Q iteration - first experiences with a data efficient neural reinforcement learning method. *Machine Learning: ECML* 2005;.
17. Porta, J.M., Vlassis, N., Spaan, M.T., Poupart, P.. Point-based value iteration for continuous POMDPs. *The Journal of Machine Learning Research* 2006;.
18. ISO24617-2, . Language resource management-Semantic annotation frame work (SemAF), Part2: Dialogue acts. ISO; 2010.
19. Schatzmann, J., Georgila, K., Young, S.. Quantitative evaluation of user simulation techniques for spoken dialogue systems. In: *Proceedings of SIGDIAL*. 2005:.
20. Meguro, T., Minami, Y., Higashinaka, R., Dohsaka, K.. Wizard of Oz evaluation of listening-oriented dialogue control using POMDP. *Proceedings of ASRU* 2011;.
21. Terrell, A., Mutlu, B.. A regression-based approach to modeling addressee backchannels. *Proceedings of the 13th Annual Meeting of SIGDIAL* 2012;.
22. Breiman, L.. Bagging predictors. *Machine Learning* 1996;.
23. Weka 3: Data Mining Software in Java, . http://www.cs.waikato.ac.nz/ml/weka/. 2009.
24. Kudo, T., Yamamoto, K., Matsumoto, Y.. Applying conditional random fields to Japanese morphological analysis. *Proceedings of EMNLP* 2004;.
25. Lee, C., Jung, S., Kim, S., Lee, G.G.. Example-based dialog modeling for practical multi-domain dialog system. *Speech Communication* 2009;.
26. Schaul, T., Bayer, J., Wierstra, D., Sun, Y., Felder, M., Sehnke, F., Rückstieß, T., Schmidhuber, J.. Pybrain. *The Journal of Machine Learning Research* 2010;.
27. Ng, A.Y., Harada, D., Russel, S.. Policy invariance under reward transformations: Theory and application to reward shaping. *Proceedings of the 16th International Conference on Machine Learning* 1999;.
28. Asri, L.E., Laroche, R., Pietquin, O.. Reward shaping for statistical optimisation of dialogue management. *Proceedings of the International Conference on Statistical Language and Speech Processing* 2013;.
29. Meguro, T., Higashinaka, R., Minami, Y., Dohsaka, K.. Controlling listening-oriented dialogue using partially observable Markov decision processes. *Proceedings of COLING* 2010;.
30. Misu, T., Georgila, K., Leuski, A., Traum, D.. Reinforcement learning of question-answering dialogue policies for virtual museum guides. *Proceedings of the 13th Annual Meeting of SigDial* 2012;.
31. Bench-Capon, T.J.M.. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation* 2003;.
32. Carenini, G., Moore, J.D.. Generating and evaluating evaluative arguments. *Artificial Intelligence* 2006;.
33. Nouri, E., Park, S., Scherer, S., Gratch, J., Carnevale, P., Morency, L.P., Traum, D.. Prediction of strategy and outcome as negotiation unfolds by using basic verbal and behavioral features. *Proceedings of INTERSPEECH* 2013;.

Appendix

Table 13: Original Japanese transcription of Table 12.

Speaker	Transcription	Fra	GPF
System	Aのカメラは、Bのカメラより、シャッターのスピードが速いです。 例えば、この顔を撮ったのに、別の顔が撮れてるっていうような、時です。 もし、シャッタースピードが速いのであれば、一瞬も逃さずに撮ることが、出来ます。	Pos A	INFORM
System	Aのカメラと他のカメラの大きな違いはセンサーです。 デジタルカメラはセンサーが命といわれています。 Aのカメラのセンサーは一眼と同じセンサーを使ってるんですね。	Pos A	INFORM
System	なお、Aのカメラの大きさは他のカメラと一緒です。	None	INFORM
System	Aのカメラはポケットに入る大きさで一眼並みの性能を持っています。	Pos A	INFORM
System	Aのカメラは価格コムとかでも一番人気の機種です	Pos A	INFORM
System	Eのカメラは軽いんです。	Pos E	INFORM
System	しかし、Aのカメラが今一番人気があります。	Pos A	INFORM
System	Aは画素数が2千90万画素でズームが3.6倍です。 Aのカメラは非常に大きなセンサーを使ってますので、細やかに画像を速く処理出来ます。	Pos A	INFORM
System	いかがでしょうか？	None	RTURN
User	AはEよりも重いんですか？	None	PROPQ
System	遠いところと近いは、どっちをよく撮りますか？	None	PROPQ
System	長い間、同じカメラを使っていると、より高性能なカメラが欲しくなってくると思います。 Aのカメラはそういうことが多分少ないと思います。 Aのカメラの性能積んでる他のカメラは、なかなかないです。	Pos A	INFORM
System	一度、色々試していただいて、決めてもらおうと思うんですけど、 Aのカメラが、断然人気なんですね。	Pos A	INFORM
System	いかがでしょうか？	None	RTURN
User	AとEどちらが安いですか？	None	PROPQ
System	始め安いカメラ買って、また買い直したら、その分お金かかかりますよね？	None	QUESTION
System	一眼レフは、10万20万が、相場です。 この相場を考えると、Aのカメラは、高くありません。	Pos A	INFORM
System	いかがでしょうか？	None	RTURN
User	Aがいいと思います。	None	INFORM
	⟨Dialogue is closed by the user ⟩		

Table 14: The summary of one dialogue in the corpus (translated from Japanese)

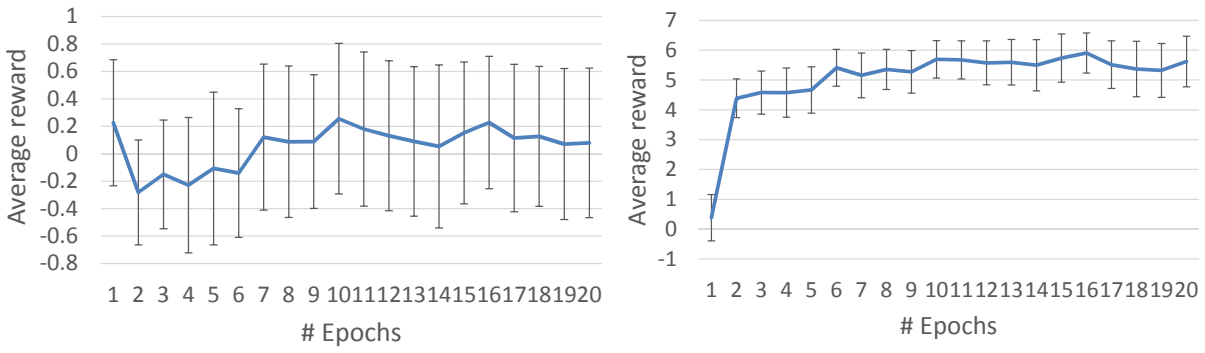
Speaker	Transcription	GPF Tag
Customer	Hello.	Inform
Customer	I'm looking for a camera for traveling. Do you have any recommendations?	PropQ
Salesperson	What kind of pictures do you want to take?	SetQ
Customer	Well, I'm the member of a tennis club, and want to take a picture of landscapes or tennis.	Answer
salesperson	O.K. You want a camera which can take both far and near. Don't you?	PropQ
Salesperson	Well, have you used a camera before?	PropQ
Customer	I have used a digital camera. But the camera was cheap and low resolution.	Answer
Salesperson	I see. I see. Camera A is a high resolution camera. A has extremely good resolution compared with other cameras. Although this camera does not have a strong zoom, its sensor is is almost the same as a single-lens camera.	Inform
Customer	I see.	Inform
Salesperson	For a single lens camera, buying only the lens can cost 100 thousand yen. Compared to this, this camera is a bargain.	Inform
Customer	Ah, I see.	Inform
Customer	But, it's a little expensive. right?	PropQ
Customer	Well, I think, camera B is good at price.	Inform
Salesperson	Hahaha, yes, camera B is reasonably priced.	Answer
Salesperson	But its performance is low compared with camera A.	Inform
Customer	If I use the two cameras will I be able to tell the difference?	PropQ
Salesperson	Once you compare the pictures taken by these cameras, you will understand the difference immediately. The picture itself is very high quality. But, camera B and E are lower resolution, and the picture is a little bit lower quality.	Answer
Customer	Is there also difference in normal size pictures?	PropQ
Salesperson	Yes, whether the picture is small or large, there is a difference	Answer
Customer	Considering A has single-lens level performance, it is surely reasonable.	Inform
Salesperson	I think so too.	Inform
Salesperson	The general price of a single-lens is about 100 or 200 thousand yen. Considering these prices, camera A is a good choice.	Inform
Customer	Certainly, I'm interested in this camera.	Inform
Salesperson	Considering its performance, it is a bargain.	Inform
Customer	I think I'll go home, compare the pictures, and think a little more.	Commisive
Salesperson	I see. Thank you.	Directive

Table 15: Original Japanese transcription of Table 14.

Speaker	Transcription	GPF Tag
Customer	あ、どうも、こんにちは	Inform
Customer	旅行用に使うカメラをちょっと探しに来たんですけど、どういったカメラがありますかね	PropQ
Salesperson	主にどういったものをとりますか	SetQ
Customer	えーっと一応サークルでテニスをやってまして その時にそのーやっぱり旅行行ったときの風景の写真だったりテニスをやってるときの 写真とかをとろうかなと思ってるんですけど	Answer
Salesperson	あー近いところもとれるし、遠いところもとれるっていう感じですかね	PropQ
Salesperson	えーっと今までなんか使ってるとかありますか	PropQ
Customer	今までは一応デジカメとかを使ったんですけど ちょっと安めのカメラを使っててなんかあんまりうまく撮れないなとかそのーなんかずれが大きかったりとか、 後画質がよくないなと思ってちょっといいのを買いない買いたいと思ってちょっと来たんですけど	Answer
Salesperson	なるほどなるほどそうですね えーっとまあ画質うんぬんでちょっと話さしてもらいますと このカメラ Aってのがあるんですけどね，こちらの方ですと画素数がかなり高めの設定になっております。 まあ他のものに比べますとちょっとまあ群を抜いてあの画素数が高めなので， まあのただ弱いところといいますとえーちょっとまあ望遠とかに関しては弱い部分もあるんですけども レンズの性能後それと内臓されてるえーセンサーこれがねちょっと大型のもの使ってまして、 えー実際ですねこのカメラに関してはあの一眼レフと同じような性能です。	Inform
Customer	あそうなんですか	Inform
Salesperson	はいあのまあちょっと一眼レフって聞くとね， やっぱりレンズだけでももう 10 万ぐらいするとかけっこうそうゆうイメージがあると思うんですけども， そのレンズを買うこと思ったら、まあこの本体自体はもうかなりまあ値段的には安めに設定されているので， そのレンズがもう一本二本買うぐらいであればこのカメラ一台で間に合うかなというところがありますね	Inform
Customer	あーそうなんですね。	Inform
Customer	でもちょっと高いですよね	PropQ
Customer	あー例えばあの B とかいいカメラってのがけっこう値段的にはいいのかなと思ってたんですけど	Inform
Salesperson	ははははそうですね値段的にね B とかいいカメラに関してはちょっと安めの設定になってますけどもね	Answer
Salesperson	あのーまあカメラ A に比べますとそうですねやっぱりちょっと性能の方も落ちるので	Inform
Customer	あーけっこうな差っていうのは使ってて感じるもんですかね	PropQ
Salesperson	そうですね実際ねこの仕上がりの写真なんですけども、 あのホームページの方とかでも発表されてるんですけど，撮り比べた写真とかをね色々見比べてみますと， やはり画素数の違いってゆうのがやっぱりもう出てきますんで、かなり絵自体はすごくきれいに撮れますね. やっぱりこのカメラ B とか E になってくるとね画素数が落ちるので， どうしてもカメラ A に比べるとやっぱりちょっと仕上がりがちょっと弱いかなっていうところがあります	Answer
Customer	あーこの画素数の差っていうのは普通に一般的なサイズの写真にしたときでも， けっこう差っていうのは出たりするんですかね	PropQ
Salesperson	差はやっぱり出ますね. ちっちゃめの写真から大きく引き伸ばした時までっていうのを， 比べていただいても明らかに差がでます	Answer
Customer	確かに一眼レフの能力があるって考えると安いですね	Inform
Salesperson	そうなんですよね	Inform
Salesperson	やっぱり一眼レフってゆうとね 10 万 20 万とかゆうのがもう相場なので， それをね考えはるとやっぱりこれはこうけっこうおすすめってとこですね	Inform
Customer	確かに少しいいカメラっていうのも興味があるんで	Inform
Salesperson	性能の割にはほんとお値段的にはお手頃なので、一度買っていただいたらほんと損はないカメラだと思いますよ	Inform
Customer	家に帰って写真とか見比べてたりしてみてもう一度検討してみたいと思います	Commisive
Salesperson	あわかりました． はいありがとうございます	Directive

Table 16: Average number of turns of each policies in dialogue with simulators (Section 6.1.1), and with real users (Section 6.2). Note that, in the simulator case, dialogue is closed if the total number of turns reaches 50.

	Random	NoFraming	Framing	Human
Average number of turns in dialogue with simulators	8.2	50.0	50.0	
Average number of turns in dialogue with real users	5.2	38.8	27.3	18.8



(a) (b)

Figure 9: Number of epochs (i.e. number of parameter updates) v.s. average reward of 20 learnt policies. Policies in (a) are the case where the reward (Equation (4)) is given only when dialogue is closed, and policies in (b) are the case where the reward is given at every turn. Error bars represents 95% confidence intervals.