Reinforcement Learning of Cooperative Persuasive Dialogue Policies using Framing

Takuya Hiraoka, Graham Neubig, Sakriani Sakti, Tomoki Toda, Satoshi Nakamura
Nara Institute of Science and Technology (NAIST), Nara, Japan
{takuya-h,neubig,ssakti,tomoki,s-nakamura}@is.naist.jp

Abstract

In this paper, we apply reinforcement learning for automatically learning cooperative persuasive dialogue system policies using framing, the use of emotionally charged statements common in persuasive dialogue between humans. In order to apply reinforcement learning, we describe a method to construct user simulators and reward functions specifically tailored to persuasive dialogue based on a corpus of persuasive dialogues between human interlocutors. Then, 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 applying reinforcement learning is effective for construction of cooperative persuasive dialogue systems which use framing.

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 (Georgila and Traum, 2011; Georgila, 2013; Paruchuri et al., 2009; Heeman, 2009; Mazzotta and de Rosis, 2006; Mazzotta et al., 2007; Nguyen et al., 2007; Guerini et al., 2003). We concern ourselves with *cooperative* persuasive dialogue systems (Hiraoka et al., 2013), which try to satisfy both the user and system goals. For these types of systems, creating a system policy that both has persuasive power and is able to ensure that the user is satisfied is the key to the system's success.

In recent years, reinforcement learning has gained much attention in the dialogue research community as an approach for automatically learning optimal dialogue policies. The most popular framework for reinforcement learning in dialogue models is based on Markov decision processes (MDP) and partially observable Markov decision processes (POMDP). In these frameworks, the system gets a reward representing the degree of success of the dialogue. Reinforcement learning enables the system to learn a policy maximizing the reward. Traditional reinforcement learning requires thousands of dialogues, which are difficult to collect with real users. Therefore, a user simulator which simulates the behavior of real users is used for generating training dialogues. Most research in reinforcement learning for dialogue system policies has been done in slot-filling dialogue, where the system elicits information required to provide appropriate services for the user (Levin et al., 2000; Williams and Young, 2007).

There is also ongoing research on applying reinforcement learning to persuasion and negotiation dialogues, which are different from slot-filling dialogue (Georgila and Traum, 2011; Georgila, 2013; Paruchuri et al., 2009; Heeman, 2009). In slot-filling dialogue, the system is required to perform the dialogue to achieve the user goal, eliciting some information from a user to provide an appropriate service. A reward corresponding to the achievement of the user's goal is given to the system. In contrast, in persuasive dialogue, the system convinces the user to take some action achieving the system goal. Thus, in this setting, reward corresponding to the achievement of both the user's and the system's goal is given to the system. The importance of each goal will vary depending on the use case of the system. For

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1706–1717, Dublin, Ireland, August 23-29 2014.

example, a selfish system could be rewarded with an emphasis on only achievement of the system goal, and a cooperative system could be rewarded with an emphasis on achievement of both of the goals. In addition, negotiation dialogue could be considered as a kind of the persuasive dialogue where the user also tries to convince the system to achieve the user's goal.

In this paper, our research purpose is learning better policies for cooperative persuasive dialogue systems using *framing*. We focus on learning a policy that tries to satisfy both the user and system goals. In particular, two elements in this work set it apart from previous works:

- We introduce framing (Irwin et al., 2013), which is known to be important for persuasion and a key concept of this paper, as a system action. Framing uses emotionally charged words 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 as a system action.
- We use a human-to-human persuasive dialogue corpus of Hiraoka et al. (2014) 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.

To achieve our research purpose, we construct a POMDP where the reward function and user simulator are learned from a corpus of human persuasive dialogue. We define system actions based on framing and general dialogue acts. In addition, the system dialogue state (namely, belief state) is defined for tracking the system's rewards. Then, we evaluate the effect of framing and learning a system policy. Experimental evaluation is done through a user simulator and real users.

2 Reinforcement learning

Reinforcement learning is a machine learning technique for learning a system policy. The policy is a mapping function from a dialogue state to a particular system action. In reinforcement learning, the policy is learned by maximizing the reward function. Reinforcement learning is often applied to models based on the framework of MDP or 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 distribution over states called a belief state b, thus $b(s_i)$ is the probability of being in state s_i, with initial belief state b_0. When the system performs an action $\alpha_i \in A$ based on b, following a policy $\pi : S \rightarrow A$, 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 Q-function, $Q^\pi : S \times A \rightarrow \Re$.

In this framework, it is critical to be able to learn a good policy function. In order to do so, we use Neural fitted Q Iteration (Riedmiller, 2005) 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 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[1].

3 Persuasive dialogue corpus

In this section, we give a brief overview of Hiraoka et al. (2014)'s persuasive dialogue corpus between human participants that we will use to estimate the models described in later sections.

[1] In a preliminary experiment, we found that Neural fitted Q Iteration had high performance compared to using the linear approximation of the Q-function in this domain.

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?	PROPQ
Sales	Yes, we have camera B.	ANSWER
Sales	Did you already take a look at it somewhere?	PROPQ
Cust	Yes. On the Internet.	ANSWER
Sales	It is very nice. Don't you think?	PROPQ
Cust	Yes, that's right, yes.	INFORM

Table 2: Sytem and user's GPF tags

Inform	Answer	Question	PropQ
SetQ	Commisive	Directive	

Table 3: An example of positive framing

(Camera A is) able to achieve performance of comparable single-lens cameras and can fit in your pocket, this is a point.

3.1 Outline of persuasive dialogue corpus

As a typical example of persuasive dialogue, the corpus consists of dialogues between a salesperson (persuader) and customer (persuadee). The salesperson attempts to convince the customer to purchase a particular product (decision) from a number of alternatives (decision candidates). This type of dialogue is defined 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.

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 of the cameras, 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 dialogues. 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.

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. Further examples are shown in Table 8 in the appendix.

3.2 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 the general-purpose functions (GPF) defined by the ISO international standard for dialogue act annotation (ISO24617-2, 2010). 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 uses emotionally charged words to explain particular alternatives. It has been suggested that humans generally evaluate decision candidates by selecting based on several determinants weighted by the user's preference, and that framing is an effective way of increasing persuasive power. This corpus focuses on negative/positive framing (Irwin et al., 2013; Mazzotta and de Rosis, 2006), with negative framing using negative words and positive framing using positive words.

In the corpus, framing is 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 represents whether the framings contains a reference to the persuadees preferred determinant (for example, the performance or price of a camera), taking the value TRUE if contained, and FALSE if not contained. The user's preferred determinant is annotated based on the results of a questionnaire.

Table 3 shows an example of 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

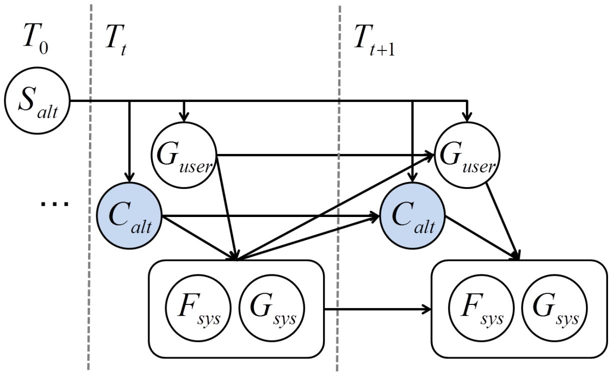

Figure 1: Dynamic Bayesian network of the user simulator. Each node represents a variable, and each edge represents a probabilistic dependency. The system cannot observe the shaded variables.

not contain any description of price. Thus, r=NO is annotated. Further examples of positive and negative framing are shown in Tables 9 and 10 in the appendix.

In this paper, we re-perform annotation of the framing tags and evaluate inter-annotator agreement, which is slightly improved from Hiraoka et al. (2014). 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 annotated to all utterances chosen by the two annotators. The inter-annotator agreement of framing polarity is 96.9% (kappa=0.903).

4 User simulator

In this section, we describe a statistical dialogue model for the user (customer in Section 3). This model 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 by 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 clerk'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)[2]. 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 5.1. Specifically, GPF and CPD are used for calculating naturalness and persuasion 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^{t+1}_{user} and CPD C^{t+1}_{alt} at turn $t+1$ are calculated by the following equations.

$$P(G^{t+1}_{user}|G^t_{user}, F^t_{sys}, G^t_{sys}, S_{alt}) \quad (1)$$

$$P(C^{t+1}_{alt}|C^t_{alt}, F^t_{sys}, G^t_{sys}, S_{alt}) \quad (2)$$

G^t_{sys} represents the system GPF at time t. F^t_{sys} represents the system framing at t. These two variables correspond to system actions, and are explained in Section 5.2. G^t_{user} represents the user's GPF at t. C^t_{alt} represents the CPD at t. S_{alt} represents the users's original evaluation of the alternatives. In our

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

research, this is the camera that the user selected as a preferred camera at the beginning of the dialogue[3]. 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. In addition, we use logistic regression for learning Equations (1) and (2).

5 Learning cooperative persuasion policies

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

5.1 Reward

We follow Hiraoka et al. (2014) in defining a reward function according to three factors: user satisfaction, system persuasion success, and naturalness. As described in Section 1, we focus on developing cooperative persuasive dialogue systems. Therefore, the system must perform dialogue to achieve both the system and user goals. In our research, we define three elements of the reward function as follows:

Satisfaction 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 The system goal is represented by persuasion success. Persuasion success represents whether the persuadee finally chooses the persuasive target (in this paper, camera A) at the end of the dialogue. Persuasion success takes the value SUCCESS when the customer decides to purchase the persuasive target at the end of dialogue, and FAILURE otherwise.

Naturalness In addition, we use naturalness as one of the rewards. This factor is known to enhance the learned policy performance for real users (Meguro et al., 2011).

The reward at each turn t is calculated with the following equation[4].

$$r_t \quad = \quad (Sat^t_{user} + PS^t_{sys} + N^t)/3 \tag{3}$$

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 persuasion 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.

$$N_t \quad = \quad P(F^t_{sys}, G^t_{sys}, G^t_{user} | F^{t-1}_{sys}, G^{t-1}_{sys}, G^{t-1}_{user}) \tag{4}$$

In our research, Sat and PS are calculated with a predictive model constructed from the human persuasion dialogue corpus described in Section 3. In constructing these predictive models, the persuasion results (i.e. persuasion success and persuadee's satisfaction) at the end of dialogue are given as the supervisory signal, and the dialogue features in Table 4 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 4 shows all features used for reward calculation at each turn[5]. Note that, for the calculating TOTAL TIME, average speaking time corresponding to speakers and dialogue acts is added at each turn.

[3] Preliminary experiments indicated that the user behaved differently depending on the first selection of the camera, thus we introduce this variable to the user simulator.

[4] We also optimized the policy in the case where the reward (Equation (3)) is given only when dialogue is closed. However, the convergence of the learning was much longer, and the performance was relatively bad.

[5] Originally, there are more dialogue features for the predictive model. However as in previous research, we choose significant dialogue features by step-wise feature selection (Terrell and Bilge, 2012).

Table 4: Features for calculating reward. These features are also used as the system belief state.

Sat_{user}	Frequency of system commisive
	Frequency of system question
PS_{sys}	Total time
	C_{alt} (for each 6 cameras)
	S_{alt} (for each 6 cameras)
N	System and user current GPF
	System and user previous GPF
	System framing

Table 5: 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 6: 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>	

5.2 Action

The system's action $\langle F_{sys}, G_{sys}\rangle$ is a framing/GPF pair. These pairs represent the dialogue act of the salesperson, and are required for reward calculation (Section 5.1). There are 11 types of framing (Table 5), 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 pairs for which the likelihood is below 0.005[6]. As a result, the 13 pairs shown in Table 6 remained[7]. We use these pairs as the system actions.

5.3 Belief state

The current system belief state is represented by the features used for reward calculation (Table 4) and the reward calculated at 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 equation.

$$P(\hat{C_{alt}^{t+1}} | \hat{C_{alt}^{t}}, F_{sys}^{t}, G_{sys}^{t}, S_{alt}) \tag{5}$$

where $\hat{C_{alt}^{t+1}}$ represents the estimated CPD at $t+1$. $\hat{C_{alt}^{t}}$ represents the estimated CPD at t. The other variables are the same as those in Equation (2). In contrast, we assume that the system can observe G_{user} and S_{alt}. G_{user} is not usually observable because traditional dialogue systems have automatic speech recognition/Spoken language understanding errors. However, in this work, we use Wizard of Oz in place of automatic speech recognition/Spoken language understanding (Section 6.2). Thus, we can ignore these factors[8].

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 and real users.

[6] 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 (Georgila, 2013).

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

[8] In addition to this reason, the G_{user} is not so essential to our research (GPF is general dialogue act), and we want to focus the CPD. This is the other reason that we assume that G_{user} is observable.

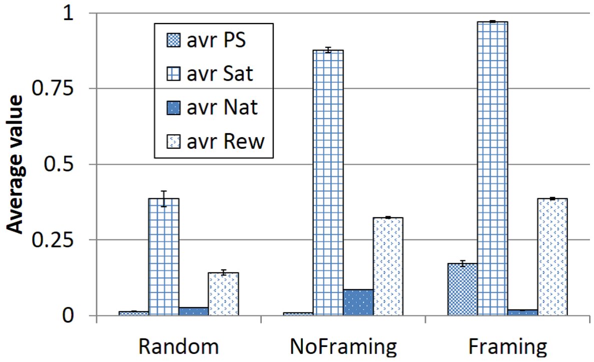

Figure 2: Average reward of each system. Error bars represents 95% confidence intervals. Rew represents the reward, Sat represents the user satisfaction, PS represents persuasion success, and Nat represents naturalness.

6.1 Policy learning and evaluation using the user simulator

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 5.2. 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 5.2 including framing. The policy is also a greedy policy.

For learning the policy, we use Neural fitted Q Iteration (Section 2). For applying Neural fitted Q Iteration, we use the Pybrain library (Schaul et al., 2010). We set the discount factor γ of learning 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 50 dialogues as one epoch, and update the parameters of the neural network at each epoch. Learning is finished when number of epochs reaches 200 (10000 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.

Experimental results (Figure 2) 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 better achieve user satisfaction than Random. On the other hand, only Framing is able to achieve better persuasion success than Random. This result indicates that framing is effective for persuasive success. In contrast, naturalness of Framing is not improved from Random. One of the reasons for this is that variance of Nat is smaller than those of the other factors, and the optimization algorithm favored the other two factors which had a higher variance.

6.2 Real user evaluation based on Wizard of Oz

To test whether the gains shown on the user simulator will carry over to an actual dialogue scenario, we perform an experiment with real human users. In addition to the policies described in Section 6.1, 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.

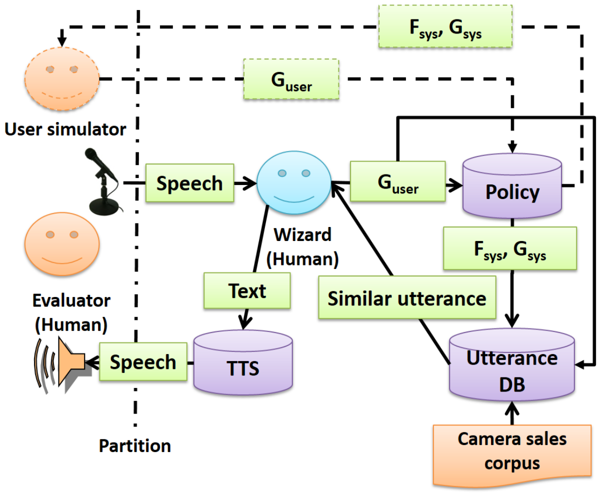

Figure 3: The experimental environment based on Wizard of Oz. The rectangle represents information, and the cylinder represents a system module. The information flow (dashed line) in the experiment through the user simulator is also shown for comparison.

Experimental evaluation is conducted, based on the Wizard of Oz framework. In the experiment, the wizard plays the salesperson, and the evaluator plays the customer. Dialogue is performed between the wizard and the evaluator. The wizard and evaluator are divided by a partition, and the evaluator cannot see or detect what the wizard is doing. The evaluator selects his/her preferred camera from the catalog before starting evaluation. Then, the evaluator starts the dialogue with the wizard who is obeying one of the policies (Figure 3). In particular, dialogue between wizard and evaluator proceeds based on the following steps.

1. The evaluator talks to the wizard using the mic. In this step, the evaluators can close the dialogue if they want.
2. The wizard listens to the evaluator's utterance, translating the utterance into the appropriate G_{user}. Then, the wizard inputs G_{user} to the policy module.
3. The policy module decides action sequences (F_{sys}, G_{sys}) based on G_{user}, then outputs the action to the utterance database module. This module is constructed from the camera sales corpus (Section 3).
4. The utterance database module searches for similar sentences that match the history of input actions and G_{user} so far, then outputs the top 6 similar utterances to the wizard.
5. The wizard generates the system utterance (Text) using the retrieved sentences. The wizard selects one sentence which best matches the context[9]. If the wizard determines the sentence is hard to understand, the wizard can correct the sentence to be more natural.
6. The wizard inputs the system utterance to text-to-speech, then waits for the next evaluator utterance (back to step 1).

Finally, the evaluator answers the following questionnaire for calculating the evaluation measures in Section 5.1.

Satisfaction The evaluator'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 customer finally wants to buy.

We use SofTalk (cncc, 2010) as text-to-speech software.

Evaluation criteria are basically same to those of previous section (described in Section 5.1). Note that in the previous section, Sat_{user} and PS_{sys} are estimated from the simulated dialogue. In contrast to the previous section, Sat_{user} and PS_{sys} are calculated from the result of the real user's questionnaire

[9] Note that the wizard is not allowed to create the utterance with complete freedom, and selects an utterance from the utterance database even when Human policy is used.

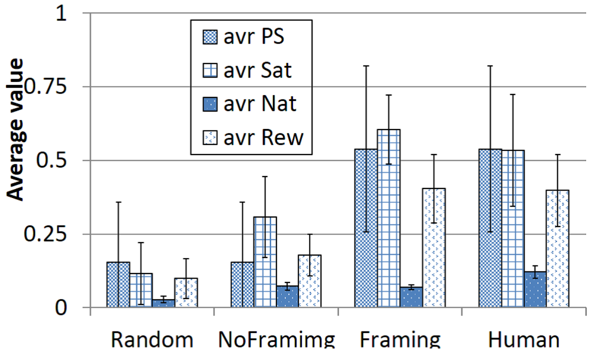

Figure 4: Evaluation results for real users. Error bars represent 95% confidence intervals. Rew represents the reward, Sat represents the user satisfaction, PS represents persuasion success, and Nat represents naturalness.

Table 7: Part of a dialogue between Framing and an evaluator (translated from Japanese)

Speaker	Transcription	Fra	GPF
Wiz	Which pictures do you want to take? Far or near?	None	QUESTION
Wiz	Camera B has 20x zoom, and this is good.	Pos B	ANSWER
Wiz	How about it?		RELEASET
Eva	I think B sounds good.		ANSWER
Wiz	Yes, B is popular with zoom,	Pos B	INFORM
Wiz	But, A has extremely good performance. Camera A has almost the same parts as a single lens camera, and is more reasonably priced than a single lens-camera.	Pos A	ANSWER
Wiz	How about it?		RELEASET

(described in the previous paragraph)[10] based on the definition of Sat_{user} and Sat_{user} in Section 6.1. The naturalness is automatically calculated by the system, in the same manner as described in the previous section. Finally, reward is calculated considering Sat_{user}, PS_{sys} and naturalness according to Equation 3.

Participants consist of 13 evaluators (3 female, 10 male) and one wizard. Evaluators perform one dialogue with the wizard obeying each policy (a total of 4 dialogues) in random order.

Experimental results (Figure 4) indicate that framing is effective in persuasive dialogues with real users, and that the reward of Framing is higher than NoFraming and Random, and almost equal to Human. In addition, the score of NoFraming is almost equal to Random. This indicates that despite the fact that it performed relatively well in the simulation experiment, NoFraming is not an effective policy for real users. In addition, the score of NoFraming is lower than the score given by the user simulator. In particular, persuasion success is drastically decreased. This indicates that framing is important for persuasion.

We can see that some features in human persuasive dialogue appear in the dialogue between users and the wizard using the Framing policy. An example of a typical dialogue of Framing is shown in Table 7. The first feature is that the system also recommends camera B when the system does positive framing of camera A, which is the persuasive target. This feature was found by Hiraoka et al. (2014) to be an indicator of persuasion success in the camera sales corpus. The second feature is that the system asks the user about the user's profile at the first stage of the dialogue. This feature is often found when user satisfaction is high. The second feature also appeared in the dialogue with NoFraming. However, NoFraming does not use framing, and asks the user to make a decision (DIRECTIVE). An example utterance from the DIRECTIVE class is "Please, decide (which camera you want to buy) after seeing the catalog".

Considering the evaluation result of the previous section, we can see that Sat and PS differ between the user simulator and the real users ($p < .05$). While the general trend of showing improvements for

[10] Note that, though systems estimate the satisfaction and evaluator's decision at each turn for the belief state, the human evaluator answers the questionnaire only when the dialogue is closed.

satisfaction and persuasive success is identical in Figures 2 and 4, the systems are given excessively high Sat in simulation. In addition, systems (especially Framing) 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 (Hiraoka et al., 2014). In this paper, dialogue features for the predictive model are calculated at each turn. In addition, persuasion 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 persuasion success and user satisfaction at each turn. Another reason is that the simulator is not sufficiently accurate to use for reflecting real user's behavior. Compared to other works (Meguro et al., 2010; Misu et al., 2012), we are using a relatively small sized corpus for training the user simulator. Therefore, the user simulator cannot be trained to accurately imitate real user behavior. Improving the user simulator is an important challenge for future work.

7 Related work

There are a number of related works that apply reinforcement learning to persuasion and negotiation dialogue. Georgila and Traum (2011) 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 are assumed as an example of negotiation dialogue. In addition, Georgila (2013) 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 (2009) models negotiation dialogue, assuming a furniture layout task, and Paruchuri et al. (2009) 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 (Irwin et al., 2013). By considering framing, the system has the potential to be more persuasive. While there is one previous example of persuasive dialogue using framing (Mazzotta et al., 2007), 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. (2014) for estimating persuasion 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

We apply reinforcement learning for learning cooperative persuasive dialogue system policies using framing. In order to apply reinforcement learning, a user simulator and reward function is constructed based on a human persuasive dialogue corpus. Then, we evaluate the learned policy and effect of framing using 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 construct a fully automatic persuasive dialogue system using framing. In this research, automatic speech recognition, spoken language understanding and natural language generation are performed by a human Wizard. We plan to implement these modules and evaluate system performance. In addition, in this research, corpus collection and evaluation are done in a role-playing situation. Therefore, we plan to evaluate the system policies in a more realistic situation. We also plan to consider non-verbal information (Nouri et al., 2013) for estimating persuasive success and user satisfaction.

References

cncc. 2010. SofTalk. http://www35.atwiki.jp/softalk/.

Kallirroi Georgila and David Traum. 2011. Reinforcement learning of argumentation dialogue policies in negotiation. *Proceedings of INTERSPEEECH.*

Kallirroi Georgila. 2013. Reinforcement learning of two-issue negotiation dialogue policies. *Proceedings of the SIGDIAL.*

Marco Guerini, Oliviero Stock, and Massimo Zancanaro. 2003. Persuasion model for intelligent interfaces. *Proceedings of the IJCAI Workshop on Computational Models of Natural Argument.*

Peter A. Heeman. 2009. Representing the reinforcement learning state in a negotiation dialogue. *Proceedings of ASRU.*

Takuya Hiraoka, Yuki Yamauchi, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2013. Dialogue management for leading the conversation in persuasive dialogue systems. *Proceedings of ASRU.*

Takuya Hiraoka, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Construction and analysis of a persuasive dialogue corpus. *Proceedings of IWSDS.*

Levin Irwin, Sandra L. Schneider, and Gary J. Gaeth. 2013. All frames are not created equal: A typology and critical analysis of framing effects. *Organizational behavior and human decision processes 76.2.*

ISO24617-2, 2010. *Language resource management-Semantic annotation frame work (SemAF), Part2: Dialogue acts. ISO.*

Esther Levin, Roberto Pieraccini, and Wieland Eckert. 2000. A stochastic model of human-machine interaction for learning dialog strategies. *Proceedings of ICASSP.*

Irene Mazzotta and Fiorella de Rosis. 2006. Artifices for persuading to improve eating habits. *AAAI Spring Symposium: Argumentation for Consumers of Healthcare.*

Irene Mazzotta, Fiorella de Rosis, and Valeria Carofiglio. 2007. PORTIA: a user-adapted persuasion system in the healthy-eating domain. *Intelligent Systems.*

Toyomi Meguro, Ryuichiro Higashinaka, Yasuhiro Minami, and Kohji Dohsaka. 2010. Controlling listening-oriented dialogue using partially observable Markov decision processes. *Proceedings of COLING.*

Toyomi Meguro, Yasuhiro Minami, Ryuichiro Higashinaka, and Kohji Dohsaka. 2011. Wizard of oz evaluation of listening-oriented dialogue control using pomdp. *Proceedings of ASRU.*

Teruhisa Misu, Kallirroi Georgila, Anton Leuski, and David Traum. 2012. Reinforcement learning of question-answering dialogue policies for virtual museum guides. *Proceedings of the 13th Annual Meeting of SigDial.*

Hien Nguyen, Judith Masthoff, and Pete Edwards. 2007. Persuasive effects of embodied conversational agent teams. *Proceedings of HCI.*

Elnaz Nouri, Sunghyun Park, Stefan Scherer, Jonathan Gratch, Peter Carnevale, Louis-Philippe Morency, and David Traum. 2013. Prediction of strategy and outcome as negotiation unfolds by using basic verbal and behavioral features. *Proceedings of INTERSPEECH.*

Praveen Paruchuri, Nilanjan Chakraborty, Roie Zivan, Katia Sycara, Miroslav Dudik, and Geoff Gordon. 2009. POMDP based negotiation modeling. *Proceedings of the first MICON.*

Martin Riedmiller. 2005. Neural fitted Q iteration - first experiences with a data efficient neural reinforcement learning method. *Machine Learning: ECML.*

Tom Schaul, Justin Bayer, Daan Wierstra, Yi Sun, Martin Felder, Frank Sehnke, Thomas Rückstieß, and Jürgen Schmidhuber. 2010. Pybrain. *The Journal of Machine Learning Research.*

Allison Terrell and Mutlu Bilge. 2012. A regression-based approach to modeling addressee backchannels. *Proceedings of the 13th Annual Meeting of SIGDIAL.*

Jason D. Williams and Steve Young. 2007. Scaling POMDPs for spoken dialog management. *IEEE Transactions on Audio, Speech, and Language Processing.*

Appendix

Table 8: 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
Clerk	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
Clerk	O.K. You want the camera which can take both far and near. Don't you?	PropQ
Clerk	Well, have you used a camera before?	PropQ
Customer	I have used a digital camera. But the camera was cheap and low resolution.	Answer
Clerk	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
Clerk	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
Clerk	Hahaha, yes, camera B is reasonably priced.	Answer
Clerk	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
Clerk	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
Clerk	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
Clerk	I think so too.	Inform
Clerk	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
Clerk	Considering its performance, it is a bargain.	Inform
Customer	I think I'll go home, compare the pictures, and think a little more.	Commisive
Clerk	I see. Thank you.	Directive

Table 9: Example positive framing of a salesperson's utterance $\langle a_i = \mathrm{B}, p_i = \text{POS}, r_i = \text{YES}\rangle$. In this example, the customer has indicated price as the preferred determinant.

Hahaha, yes, camera B is reasonably priced.

Table 10: Example negative framing of a salesperson's utterance $\langle a_i = \mathrm{B}, p_i = \text{NEG}, r_i = \text{NO}\rangle$. In this example, the customer has indicated price as the preferred determinant.

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).