PAPER

Hierarchical Argumentation Structure for Persuasive Argumentative Dialogue Generation

Kazuki SAKAI[†,††a)], ***Nonmember***, **Ryuichiro HIGASHINAKA**[†††], ***Member***, **Yuichiro YOSHIKAWA**[†,††], ***Nonmember***, **Hiroshi ISHIGURO**[†,††], ***Member***, ***and*** **Junji TOMITA**[†††], ***Nonmember***

SUMMARY Argumentation is a process of reaching a consensus through premises and rebuttals. If an artificial dialogue system can perform argumentation, it can improve users' decisions and ability to negotiate with the others. Previously, researchers have studied argumentative dialogue systems through a structured database regarding argumentation structure and evaluated the logical consistency of the dialogue. However, these systems could not change its response based on the user's agreement or disagreement to its last utterance. Furthermore, the persuasiveness of the generated dialogue has not been evaluated. In this study, a method is proposed to generate persuasive arguments through a hierarchical argumentation structure that considers human agreement and disagreement. Persuasiveness is evaluated through a crowd sourcing platform wherein participants' written impressions of shown dialogue texts are scored via a third person Likert scale evaluation. The proposed method was compared to the baseline method wherein argument response texts were generated without consideration of the user's agreement or disagreement. Experiment results suggest that the proposed method can generate a more persuasive dialogue than the baseline method. Further analysis implied that perceived persuasiveness was induced by evaluations of the behavior of the dialogue system, which was inherent in the hierarchical argumentation structure.

key words: *argumentation structures, argumentative dialogue, dialogue generation, persuasiveness*

1. Introduction

Argumentation is the process of reaching a consensus through premises and rebuttals [1]. It has long been studied in the fields of rhetoric and informal logic, and recently in the field of artificial intelligence [2]. Argumentation is integral in decision-making and negotiating with others. Thus, if an artificial dialogue system can aptly perform argumentation, it can assist users in improving their decision-making and negotiating.

To develop such an argumentative dialogue system, a structured database of argumentation structure is required. Argumentation structure is a graph structure; the nodes represent the statement, such as conclusions and premises; and, the edges represent the relationship between two nodes. The system can then generate a dialogue by tracking the nodes on the argumentation structures. AIFdb [3] is one of the largest argumentation structures currently available. Some studies attempted to build dialogue systems by associating a spoken sentence (utterance) to the elements of AIFdb, which would then represent a statement in the argument [4], [5]. Rach et al. reported that an argument between artificial agents whose dialogues were generated according to the hand-made argumentation structure had more logical consistency than one between humans [6]. Sakai et al. built large-scale databases to represent argumentation structures [7], creating a dialogue system to provide users with an opportunity to discuss daily topics [8], [9]. However, these systems were not equipped with a function to choose utterances according to users' responding attitudes, nor has the persuasiveness of their utterances been evaluated.

Here, a method is proposed involving a hierarchical argumentation structure [7] to accommodate human agreement and disagreement and generate persuasive arguments. In the argumentation structure, the nodes represent utterances corresponding to claims, propositions, and premises. The links between nodes represent the relationship of utterances, wherein the relationship can be marked as either supportive or non-supportive. When the statement of the system is not accepted, it produces a child statement node, which is supportive and intended to convince the user. Once the user agrees with the statement, the system re-checks the user's attitude toward the statement registered in its parent node. Persuasiveness was subjectively evaluated through the visual evaluation of dialogue texts. The proposed method was then compared to the baseline method, wherein the argument response texts were generated by the same system; however, the baseline method utilizes a flat rather than hierarchical structure, which ignores user's agreement or disagreement in generating responses.

The remainder of this paper is organized as follows: In Sect. 2, related work regarding argumentation is described. In Sect. 3, the development of the argumentative dialogue system and a method to utilize hierarchy in the argumentation structure is described. In Sect. 4, the methods and findings of the experimentation are described. In Sect. 5, the results are discussed and the factors' contribution to the persuasiveness further analyzed. Finally, the work is concluded in the Sect. 6.

2. Related Work

Several studies have investigated computational models of

Manuscript received May 29, 2019.
Manuscript revised September 22, 2019.
Manuscript publicized October 30, 2019.

[†] The authors are with Osaka University, Toyonaka-shi, 560–8531 Japan.
[††] The authors are with JST ERATO, Toyonaka-shi, 560–8531 Japan.
[†††] The authors are with NTT Media Intelligence Laboratories, NTT Corporation, Yokosuka-shi, 239–0847 Japan.
a) E-mail: sakai.kazuki@irl.sys.es.osaka-u.ac.jp

DOI: 10.1587/transinf.2019EDP7147

argumentation. Toulmin proposed an argumentation structure in which a conclusion is drawn with a datum through a warrant [10]. Other computational structures have also been proposed [11]–[13]. These adopted argumentation structures represented by graphs, wherein nodes represent statements, and edges represent supporting or non-supporting relationships between two nodes. In the present study, the argumentation structure described in [12] was adopted. Recently, many studies in the field of argument mining [14] have focused on automatically extracting premises and conclusions from texts. This has been applied to various types of text, including legal documents [15], news articles [16], opinions in discussion forums [17], and various online texts [18].

Some studies developed argumentative dialogue systems through argument mining [19]–[22]. Lawrence et al. developed a debate system that utilized an argumentation structure automatically created through an argument mining technique [19]. Rakshit et al. developed an arguing bot that chooses utterances from corpora, including debates on the relevant topic; they also explored potential structures of the corpora to expedite choices [20]. Dieu-Thu et al. developed two types of argumentative dialogue agents: a retrieval-based system using long short-term memory (LSTM) and a generative model-based system using a recurrent neural network (RNN) [21]. Marumoto et al. developed a debating system regarding TV news that generates claims and reasons by extracting appropriate sentences from the internet sources [22]. However, these systems were not capable of considering a user's opinion in generating utterances, nor were the systems' utterances evaluated for the user's impression of the dialogue's persuasiveness.

Many studies regarding argumentation model the persuasion processes. Bex and Walton proposed an extended model for an argumentative dialogue system that inserts utterances to explain reasons from the past claims of the system [23]. Hunter built an artificial agent that can select abstract argumentative actions to persuade its artificial interlocutor; this relied on a probabilistic model of the agent's perception of the appearance of the premises and conclusions in the argument [24]. Thimm discussed strategies for multi-agents to select abstract argumentative actions to accomplish the argumentation-based negotiation with another artificial agent [25]. Rosenfeld and Kraus presented a methodology for persuading people through argumentative dialogues by combining theoretical argumentation modeling, machine learning, and Markovian optimization techniques [26]. Georgila and Traum built a reinforcement learning agent that can choose dialogue actions to achieve argumentation in a simulated conversation with another agent [27]. Koit developed a mechanism for argumentation-based negotiation and dialogue strategies enabling an artificial agent to influence its artificial interlocutor to accept its request [28]. Since these studies focused on modeling the argumentation mechanism in the abstract level by computer simulation, the agent's ability to provide human users with a dialogue for argumentation remains unclear.

3. Persuasive Argument Generation Method

In this section, a method is described to consider human agreement and disagreement and generate a persuasive argument with the hierarchical argumentation structure. Figure 1 shows the process of a system generating a dialogue. The user selects one of the options, agree or disagree, as the input of the system. When the user selects one option, the dialogue manager updates the currently selected node of the structure according to the option selected. Subsequently, the dialogue manager selects a next action according to the user's agreement or disagreement. The dialogue manager also updates the selected node and sends the selected action to the natural language generation (NLG) module. The NLG module generates the surface text through ad-hoc rules.

3.1 Argumentation Structures

Figure 2 shows the argumentation structure described in [7]. The argumentation structure is a simplified version of the conventional argumentation model [12] and is adapted for dialogue usage. The proposed argumentation structure accommodates exchange of opinions between two participants

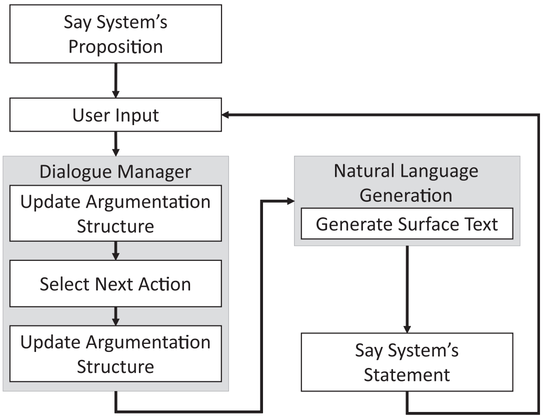


Fig. 1 Utterance generation process.

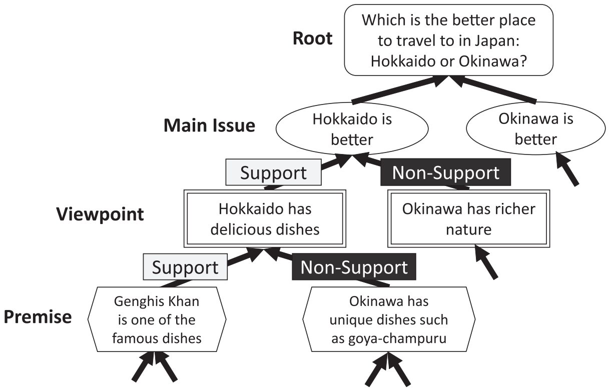

Root
Which is the better place to travel to in Japan: Hokkaido or Okinawa?
Main Issue
Hokkaido is better
Okinawa is better
Support
Non-Support
Viewpoint
Hokkaido has delicious dishes
Okinawa has richer nature
Support
Non-Support
Premise
Genghis Khan is one of the famous dishes
Okinawa has unique dishes such as goya-champuru

Fig. 2 Simplified hierarchical argumentation structure of the conventional model.

arguing various points. In the argumentation structure, the nodes represent utterances corresponding to claims, propositions, and premises. The links between nodes represent the relationship of the utterances, which are marked as either supportive or non-supportive. In this hierarchical argumentation structure, each layer in the structure serves a different purpose. The top node, or root node, represents the main proposition of the argumentation. The two nodes connected to the root node, or the main issue nodes, represent opposite stances. Connected below the main issue nodes are the viewpoint nodes. The viewpoint nodes represent conversational topics. Under each viewpoint node are premise nodes, which represent reasons regarding each topic.

3.2 Dialogue Manager

It is assumed that the development of a dialogue can be understood by tracking which node in the argumentation structure is mentioned or questioned. To handle the current state of the dialogue, three flags are defined for each node in the system; these flags represent whether the statement corresponding to the node has been mentioned, questioned, or accepted. The flag associated with each node changes according to five possible dialogue-acts.

3.2.1 Definition of Flags and Dialogue-Acts

The definitions of the flags are described as follows:

Mentioned flag (M flag) M flag can either be *True* or *False*. When the content of the node has been mentioned by the user or the system, this flag is set to *True*. The default value is *False*.

Questioned flag (Q flag) Q flag can either be *True* or *False*. When the content of the node has been questioned, this flag is set to *True*. The default value is *False*.

Accepted flag (A flag) A flag can either be *Accepted*, *Defeated*, or *Undefined*. When the content of the node has been accepted, it becomes *Accepted*. If it is rejected, it becomes *Defeated*. The default value is *Undefined*.

The dialogue manager assigns the default values of the three flags to all nodes when the system launches.

The dialogue-acts are defined as:

Assertion This dialogue-act indicates the mention of a premise.

Question This dialogue-act represents a question regarding the veracity of a premise.

Concession This dialogue-act indicates a concession of a premise. When a claim cannot be refuted, and the individual admits the claim, then it is a concession.

Retraction This dialogue-act indicates a retraction of a premise. When evidence for a challenge cannot be furnished and the individual admits it, then it is a retraction.

Other This dialogue-act indicates independent utterances of argumentation, such as greetings.

These dialogue-acts are developed in [12] and are important in performing argumentation.

3.2.2 Update Argumentation Structure

Through the use of dialogue-acts, the state of nodes in the argumentation structure can be updated. To do so, the system must first set a target node to be updated; the system focuses on the last updated node as the target node. The default target node is the root node. For the target node, the state of the node is changed according to the following rules:

- If the dialogue-act of an utterance is *Assertion*, the M flag in the target node is set to *True*.
- If the dialogue-act of an utterance is *Question*, the Q flag in the target node is set to *True*.
- If the dialogue-act of an utterance is *Concession*, the A flag in the target node is set to *Accepted*.
- If the dialogue-act of an utterance is *Retraction*, the A flag in the target node is set to *Defeated*.

3.2.3 Select Next Action

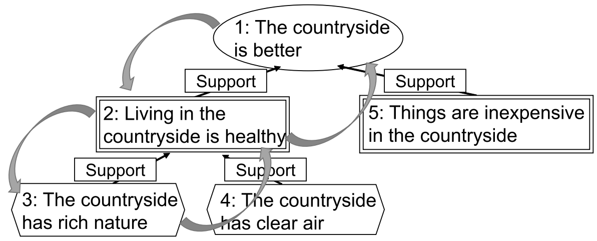
In a persuasive dialogue system, the generated statements should be perceived as orderly and logical by the user. Accordingly, when the statement of the system is not accepted, a child statement, or supportive node, is produced to convince the user. Once the user agrees with the statement, the system re-checks the user's attitude to the statement registered in the parent node.

In practice, the strategy of selecting next utterances is implemented in the following steps.

1. If the A flag of the target node is *Accepted*, the system selects its parent node, a viewpoint node, to confirm the statement. It then also selects a grandparent node, a main issue node, to let the user reconsider the user's attitude. If the target node is a viewpoint node, the system only selects its parent, main issue node.
2. If the Q flag of the target node is *True*, the system selects its unmentioned child nodes to assert its reason. If the system does not have any unmentioned child nodes, the next action candidate is nothing.
3. If the M flag of the target node is *True*, the system selects its unmentioned child nodes to assert its reason. If the system does not have any unmentioned child nodes, the next action candidate is nothing.
4. Else, the next action is to claim the target node.

If the system can determine the next action candidates, the system chooses one of them and outputs it. If the system cannot find a viable next action candidate, then the system moves from the current target node to its parent node. Then, based on the new target node, the system attempts to select the next action using the above-mentioned steps.

Figure 3 illustrates an example dialogue flow using the


1: The countryside is better
Support
Support
2: Living in the countryside is healthy
5: Things are inexpensive in the countryside
Support
Support
3: The countryside has rich nature
4: The countryside has clear air

Target Node	Speaker	Utterance	Dialogue-act	Node 1			Node 2			Node 3		
				M	Q	A	M	Q	A	M	Q	A
1	System	*The countryside is better.*	*Assertion*	→ *True*								
1	User	*I disagree*	*Question*		→ *True*							
2	System	*Really? But living in the countryside is healthy.*	*Assertion*				→ *True*					
2	User	*Not decided*	*Question*					→ *True*				
3	System	*Really? But the countryside has rich nature.*	*Assertion*							→ *True*		
3	User	*I agree*	*Concession*									→ *Accepted*
2	System	*I see. Therefore, you also agree that living in the countryside is healthy.*	*Concession*						→ *Accepted*			
1	System	*So, do you agree that the countryside is better?*	*Question*									

Fig. 3 Example dialogue flow in the proposed hierarchical method. The changes in the flags based on the dialogue-act and the transition of the target node based on the current flags are described in the table. There are three flags associated to each node, that is, Mentioned flag (M), Questioned flag (Q), and Accepted flag (A). The arrow in each column of the flags represents the changes to the new value, that is, '→ *True*' and '→ *Accepted*' indicate that the corresponding flag is changed to *True* and *Accepted*, respectively.

proposed method on a hierarchical structure consisting of five nodes. For all nodes, the initial values of M and Q flags are set to *False* while those of A flags are set to *Undef.* First, the target node is set to the node 1 placed at the top of the hierarchical structure, which corresponds to the main issue node (see the first row in the table). Because the M flag of the target node is *False*, according to the proposed strategy given in the previous paragraph, the system asserts its statement, "*The country side is better*," which corresponds to an *Assertion* dialogue-act. Consequently, the M flag is updated to *True*. Then, in this example, the user disagrees to it (see the second row in the table). In the proposed system, a user's disagreement is considered a *Question* dialogue-act because disagreeing implies that the veracity of the previous system's statement is questionable. Therefore, the user's disagreement is treated as *Question* and the Q flag is updated to *True*. Because this Q flag was set to *True*, according to the proposed strategy, the system changes the target node to node 2, which is one of its unmentioned children nodes, and then asserts its statement, "*Living in the countryside is healthy*." Consequently, it updates the M flag of the new target node to *True* (see the third row in the table). Here, in making the actual utterance, the system adds a discourse marker "*Really? But*" chosen by an ad-hoc rule to augment the cohesion of the dialogue, which will be explained in the next subsection. Until the user agrees to the system's statement, the system repeats the above processes, as shown in the fourth and fifth rows in the table of Fig. 3. Then, the user agrees to the statement as in the sixth row. In the proposed system, a user's agreement is considered a *Concession* dialogue-act because agreeing means that the user concedes the previous system's statement. Therefore, the user's agreement is regarded as *Concession* and the A flag is updated to *Accepted*. Because this A flag was set to *Accepted*, according to the proposed strategy, the system changes the target node to its viewpoint node. Then, it asserts its statement using a *Concession* dialogue-act with the discourse marker, "*I see. Therefore, you also agree that living in the countryside is healthy,*" and the A flag of the new target node is updated to *Accepted* (see the seventh row in the table). Here, since the user has conceded its child premise, the system assumes that the user also concedes its parent premise; hence the utterance corresponds to the *Concession* dialogue-act. Furthermore, the system changes the target node to the main issue node and asks the user if he/she agrees with the system's stance, "*So, do you agree that the countryside is better?*" using a *Question* dialogue-act (see the eighth row in the table). Consequently, the Q flag of the main issue node is updated to *True* (its Q flag was already set to *True*; therefore, the flag is not changed).

3.3 Natural Language Generation

The NLG module modifies the surface text according to the ad-hoc rules to make the expressions in the statement suitable to the system's dialogue-act or augment the cohesion of the dialogue. The rules were implemented by us in an ad-hoc way. First, the module obtains the utterance text from the argumentation structure, since the utterance text is associated with the corresponding node in the structure. Then, according to the selected next action, the module changes the surface text of the utterance represented in the node. For example, when the next action is to confirm and recheck the parent node, the phrase "*Therefore, you also agree that*" is concatenated to the beginning of the utterance for the confirmation node, and the phrase "*So, do you agree that*" is concatenated to the beginning of the utterance for the rechecking node.

Moreover, to augment the cohesion of the dialogue, discourse markers are included. For example, when the dialogue-act of a user's utterance is "*Concession*," the phrase "*I see.*" is added to the beginning of the system's utterance. When the dialogue-act of user's utterance is "*Question*," the phrase "*Really? But,*" is added to the beginning of the system's utterance.

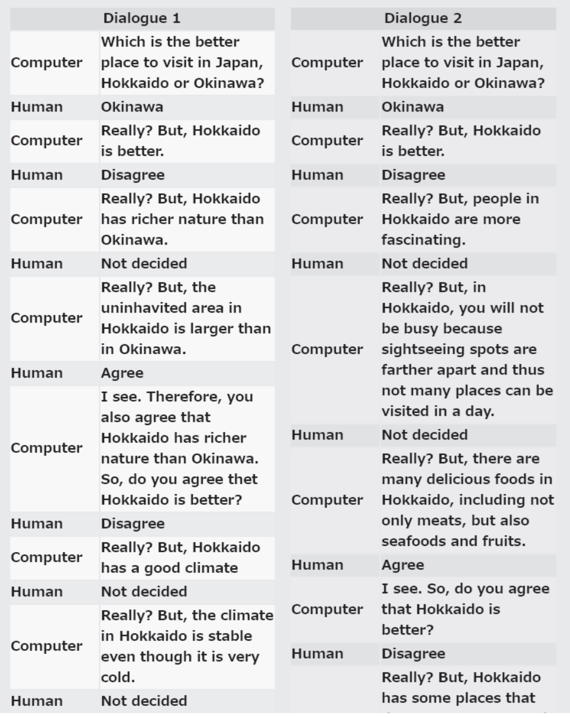
4. Evaluation

To verify the effectiveness of the proposed method, a subjective experiment was conducted. Participants read two types of argumentative dialogue texts, and then completed a questionnaire regarding their impression of each dialogue text. The persuasiveness of the dialogue was evaluated by a third person Likert scale method. Additionally, to analyze the aspects of the proposed dialogue system, multiple regression analysis were conducted. In these, the objective variable was persuasiveness. The explanatory variables were the participant's evaluations of the details of its behavior, as well as participant-side factors such as traits, opinions, or understanding of the topic argued in the given texts.

4.1 Method

4.1.1 Subjects

Two hundred adults registered to a Japanese crowdsourcing service (87 males and 113 females ranging from teens to septuagenarian) applied to this experiment. In the experiment, to collect only reliable data, we collected data only from the subjects who had high scores on their reliability for task achievement in the crowdsourcing system. Namely, we only included persons who had completed more than 95% of the task without problems so far. In addition, we did not use the data from subjects who did not pass the manipulation check question that had to be correctly answered if the subjects understood the dialogue. All subjects read a dialogue set including two dialogue texts; one was generated by the proposed method, and the other by the baseline method. Each subject read one of the ten dialogue sets, which were randomly assigned.



Dialogue 1	
Computer	Which is the better place to visit in Japan, Hokkaido or Okinawa?
Human	Okinawa
Computer	Really? But, Hokkaido is better.
Human	Disagree
Computer	Really? But, Hokkaido has richer nature than Okinawa.
Human	Not decided
Computer	Really? But, the uninhavited area in Hokkaido is larger than in Okinawa.
Human	Agree
Computer	I see. Therefore, you also agree that Hokkaido has richer nature than Okinawa. So, do you agree thet Hokkaido is better?
Human	Disagree
Computer	Really? But, Hokkaido has a good climate
Human	Not decided
Computer	Really? But, the climate in Hokkaido is stable even though it is very cold.
Human	Not decided

Dialogue 2	
Computer	Which is the better place to visit in Japan, Hokkaido or Okinawa?
Human	Okinawa
Computer	Really? But, Hokkaido is better.
Human	Disagree
Computer	Really? But, people in Hokkaido are more fascinating.
Human	Not decided
Computer	Really? But, in Hokkaido, you will not be busy because sightseeing spots are farther apart and thus not many places can be visited in a day.
Human	Not decided
Computer	Really? But, there are many delicious foods in Hokkaido, including not only meats, but also seafoods and fruits.
Human	Agree
Computer	I see. So, do you agree that Hokkaido is better?
Human	Disagree
	Really? But, Hokkaido has some places that

Fig. 4 Part of the first page of the experiment site. One dialogue was generated by the proposed method while the other was generated by the baseline method. Note that the dialogues in the experiment were shown in Japanese.

4.1.2 Apparatus

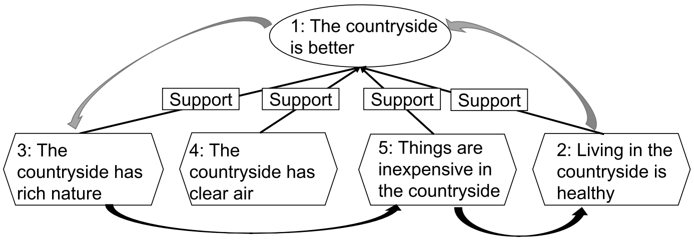
This experiment was conducted through the internet. The crowdsourcing site, a crowdsourcing web service, was used to recruit subjects and for subjects to submit completed questionnaire. The experiment site, another separate web site, provided the dialogue texts for the evaluation and questionnaire. The experiment site consisted of two pages. The first page showed the explanation of the experiment and the two dialogue texts to be evaluated. Figure 4 shows a section of the first page of the experiment site. The second page consisted of the questionnaire form.

4.1.3 Stimuli

In the experiment, two conditions were compared: hierarchical and flat conditions. The dialogue flows of each condition are shown in the following procedure:

Hierarchical Condition The system generates the dialogue by the proposed method using the hierarchical argumentation structure. Details of the generation method are described in Sect. 3.2.3.

Flat Condition The system generates the dialogue by the baseline method described in Fig. 5. As shown in the figure, this method utilizes a flat structure; the main issue node has only premise nodes, and all premise nodes have no child nodes. In experimentation, all premise


1: The countryside is better
Support
Support
Support
Support
3: The countryside has rich nature
4: The countryside has clear air
5: Things are inexpensive in the countryside
2: Living in the countryside is healthy

Target Node	Speaker	Utterance	Dialogue-act	Node 1			Node 2			Node 3			Node 5		
				M	Q	A	M	Q	A	M	Q	A	M	Q	A
1	System	*The countryside is better.*	*Assertion*	→ *True*											
1	User	*I disagree*	*Question*		→ *True*										
3	System	*Really? But the countryside has rich nature.*	*Assertion*							→ *True*					
3	User	*Not decided*	*Question*								→ *True*				
5	System	*Really? But things are inexpensive in the countryside.*	*Assertion*										→ *True*		
5	User	*I disagree*	*Question*											→ *True*	
2	System	*Really? But living in the countryside is healthy*	*Assertion*				→ *True*								
2	User	*I Agree*	*Concession*						→ *Accepted*						
1	System	*I see. So, do you agree that the countryside is better?*	*Question*												

Fig. 5 Example dialogue flow in the baseline flat method. See the notation of the table in Fig. 3.

nodes used in this flat condition were the same nodes that appeared as the viewpoint and premise nodes used in the hierarchical condition. Until the user selects "*Agree*," the system continues to state reasons by randomly selecting one node from all remaining premise nodes (e.g., the third, fifth, and seventh row in the table of Fig. 5). When the user selects "*Agree*," the system moves to the main issue node and asks if the user agrees with system's associated stance (e.g. ninth row in the table).

Notably, the information given to the participants was designed to be suitable to both the hierarchical condition and the flat condition. The premise nodes were carefully chosen such that they could suffice as premise nodes not only of the parent viewpoint node, but also of the grand-parent main issue node. The statements for the premise nodes of the viewpoint nodes in the hierarchical structure are expected to work as the direct premise of the main issue nodes. For example, "*Countryside has rich nature* (i.e., a premise node)" can be a direct premise of "*Living in the countryside is healthy* (i.e., viewpoint node)" in Fig. 3, while it can also be a direct premise of "*Living in countryside is better than city* (i.e., the main issue node)" in Fig. 5. However, some of such premise nodes do not necessarily work as the direct premises of the main issue node because the relation may not be a direct one. Therefore, the premise nodes that may not be appropriate as premises of main issue node were excluded from the structure for the flat condition in Fig. 5. Concretely, a stimulus dialogue was first tentatively created by repeating random choices of premise nodes from the premise node set. Then, if there were any premise nodes that were not appropriate as direct premises of the main issue node, these non-direct premise nodes were excluded from the premise node set. Then, the tentative dialogue was resampled until the dialogue with only premises that were appropriate as direct premises of the main issue node could be generated.

The differences between the two conditions are as follows:

(1) Choice of the node when "Agree" is not selected: in the hierarchical condition, the child node of the last mentioned node is chosen, whereas in the flat condition, another premise node is chosen.

(2) Choice of the node when "Agree" is selected: in the hierarchical condition, the viewpoint node is chosen before the main issue node is chosen, whereas in the flat condition, the main issue node is chosen.

When the user does not choose "Agree," i.e., chooses "Disagree" or "Not decided," for the viewpoint node statement in the hierarchical condition, he/she is shown its child node statement, which is a further supportive reason which he/she may then choose to accept instead. Contrastingly, when the user does not choose "Agree" for the premise node statement in the flat condition, he/she is given another premise that directly supports the main issue node, which may not necessarily be a child node of the previous premise in the hierarchical condition. For example, in Fig. 5, when the user does not choose "Agree" for "*The countryside has rich nature*," he/she is shown "*Things are inexpensive in the countryside*." It should be noted that the latter is not the child of the former, as seen in Fig. 3. Similarly, the next premise after the user disagrees is "*Living in the countryside is healthy*," which is not the child node of the disagreed premise. Difference (1) allows the dialogue to include more explicit reasons in the hierarchical condition. Furthermore, difference (2) allows the dialogue to be more logically ordered in the hierarchical condition, i.e., when the user agrees to the premise, he/she can again go through the premises, the proposed viewpoint, and the main issue, in order.

The following five topics were used:

Auto-driving Do you accept driving automobiles: yes or no?

Living In which do you prefer to live: the countryside or the city?

Sightseeing Which is the better place to visit in Japan: Hokkaido or Okinawa?

Breakfast Which is the better breakfast: bread or rice?

Theme park Which is the better theme park: Tokyo Disney Resort or Universal Studios Japan?

All five structures for each of the five topics are written in Japanese. The average number of nodes in the structures is 471.4, and the average depth of the structures is 5.3. Ten dialogue texts (two dialogues × five topics) were prepared in each condition. Each dialogue text was prepared in advance. The utterances of a simulated user to responding to the system's utterances were randomly chosen from three options: "*Agree*," "*Not decided*," and "*Disagree*." The utterances of the system were then automatically generated in response to the user's utterances. This exchange of messages was repeated 16 times. The simulated user's utterance of "*Agree*" was set to occur four times for all dialogues.

4.1.4 Procedure

On the crowdsourcing site, participants read the explanation of this experiment and then elected to participate in it. They then moved to experiment site and read the instructions shown at the top of the first page. The instructions directed the participants to read the two provided dialogue texts and to complete the ensuing questionnaire. The instructions also informed participants that the questionnaire page was shown after the dialogue page, and that the dialogues should be read carefully because the questionnaire would inquire about the content of the dialogue. Participants then began reading the two dialogue texts shown on the first page of the experiment site. Once finished reading, they pressed a button positioned at the bottom of the page. This relocated them to the second page to complete the questionnaire regarding their impression of the texts and their psychological background. After participants finished the questionnaire, they downloaded the completed form containing their responses and submitted it through the crowdsourcing site.

4.1.5 Evaluation Criteria

To evaluate the persuasiveness of the dialogue text, a questionnaire was completed by the participants regarding their impressions of the dialogues. The questionnaire had six items and was identical for both the flat and hierarchical conditions. The questionnaire included one item concerning content for a manipulation check and five items concerning user impressions for evaluation of persuasiveness. The item concerning the dialogue content asked what the system's stance was. The item concerning user's impressions of the dialogue consisted of the following five items:

Q1 Did you understand the dialogue?

Q2 Do you think the sequence of the system's statements is persuasive?

Q3 Do you think the system's statements are claimed in a logical order to support its own stance of "X" (X is either option taken by the system in each dialogue)?

Q4 Do you think the system gives explicit reasons for its own previous statement?

Q5 Do you think the system's response statements reflect the prior human statements?

A Likert scale was used to score the participates' impressions. This involved a seven-point scale that ranged from the value of 1, corresponding to "strongly disagree," to 7, corresponding to "strongly agree." The midpoint value of 4 corresponded to "undecided."

The questionnaire regarding participants' psychological background was also prepared to analyze whether the participants' impressions of the detailed aspects of the system's behavior was derived from the system's persuasiveness rather than other subject-side factors such as participant's characteristics, opinions, or understanding of the discussion topics. To measure participants' backgrounds, social skills were used, particularly KiSS-18 [29]. KiSS-18 is an influential factor to evaluate Japanese subject's impression regarding artificial systems [30]. It was assumed that a participant's social skills might affect their evaluation of the impression of the dialogue. Accordingly, another questionnaire was prepared wherein subjects answered content-specific questions regarding the proposition claimed by the system. To confirm that participants correctly understood the dialogue contents, the participants' answers were reviewed for similarity with the main claim of the system.

Table 1 Number of participants assigned to each topic.

	Auto driving	Living	Sightseeing	Breakfast	Theme park
Male	18	12	16	15	20
Female	31	19	16	24	16

4.2 Result

Of the 200 subjects who elected to participate, 187 subjects (81 males and 106 females) were used in the analysis; a subject who had trouble with the website platform and 12 subjects who failed the manipulation check were excluded. Table 1 shows the number of subjects assigned to each topic, which are used in the analysis. A significant difference in

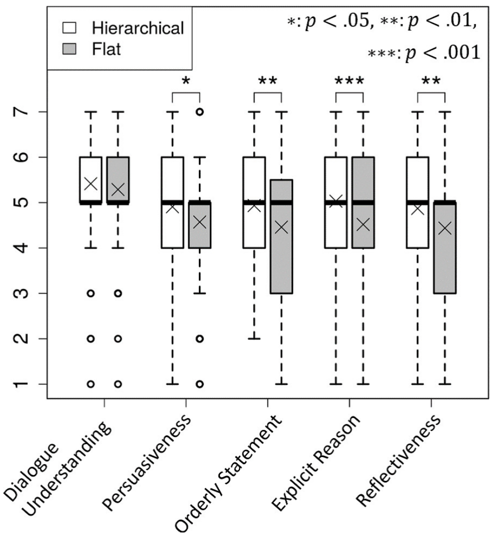


Fig. 6 Box plots of the questionnaire results. Cross marks represent the mean score.

the number of subjects among the structures and gender was not found with the χ^2-test ($\chi^2 = 4.29$, $df = 4$, $p = 0.369$).

Figure 6 shows the box plots of the questionnaire results. A paired t-test was used to compare the scores with the Bonferroni correction. The effect size also was calculated based on *Cohen's d*.

Resultingly, for Q2, "Do you think the sequence of the system's statements is persuasive?," the mean score for the hierarchical condition was significantly higher than that for the flat condition ($t = 2.63$, $p = .00892$, $d = 0.192$). For Q3, "Do you think the system's statements are claimed in a logical order to support its own stance?," the mean score for the hierarchical condition was significantly higher than that for the flat condition ($t = 3.63$, $p = .000323$, $d = 0.266$). For Q4, "Do you think the system gives explicit reason for its own previous statement?," the mean score for the hierarchical condition was significantly higher than that for the flat condition ($t = 3.86$, $p = .000132$, $d = 0.282$). For Q5, namely "Do you think the system's response statements reflect the prior human statements?," the mean score for the hierarchical condition was significantly higher than that for the flat condition ($t = 3.25$, $p = .00125$, $d = 0.238$).

5. Discussion

5.1 Validity of the Proposed Method

The results of the experiment suggest that the dialogue produced by the proposed strategy utilizing the hierarchical argumentation structure was more persuasive than that of the flat structure. In one experimental setting, the system tried to convince a user to accept the system's opinion, which conflicted with the user's current opinion. This implies that the proposed method could be used to develop a dialogue system that provides users with recommendations for more convincing strategies or conversational opportunities to convey their own stances and opinions. The current results are relatively limited because they rely on evaluations, of participants' impressions, by a third person who did not participate in the argument but observed it. Third person evaluation was adopted because of the difficulty with which participants may be able to objectively or calmly evaluate the persuasiveness; emotional or irrepressible repulsions caused by personal factors, including a participant's original stances on the current topic, artificial systems, argumentation, may affect their interpretation of the system's persuasiveness. As future work, further experiments should ask subjects to directly interact with the proposed system, while carefully controlling for the subjects' characteristics and attitudes regarding the discussed topic.

The proposed hierarchical argumentation structure generated utterances by choosing supportive child nodes. However, non-supportive nodes exist in the structure, but remain unused. By using non-supportive child nodes, it is possible for the system to generate a more persuasive argument: Wolfe et al. reported that an argument including rebuttals was more agreeable than the one without [31]. Accordingly, in future work, the next action selection method should also be improved through the use of non-supportive child nodes to allow dialogues to include rebuttals.

5.2 Effectiveness of the Proposed Method

The contribution of the dialogue's various aspects impacted users' impressions of its persuasiveness. When the human did not agree with the statement of the viewpoint node, it then mentioned one of its premise nodes. Then, when the human showed agreement to the premise, the proposed system re-checked human agreement to the current viewpoint. Since the nodes chosen in the first step were premise nodes supportive to the viewpoint, this re-checking was intended to supply further reasoning as to why the system mentioned the current viewpoint. Accordingly, as indicated by the result of Q4, this function was perceived as the system's ability to claim reasons more explicitly than it did in the flat condition. Meanwhile, this additional statement and re-checking function was employed depending on the human statement, such that the dialogue could appear to have considered and reflected the human statement. As indicated by the result of Q5, this achieved the aim for human-impacted conversational statements. Additionally, the developing dialogue maintained order in the proposed system: mentions of a premise always appeared after the human's utterance, agreement to the viewpoint was not shown while the system re-checked the user's stance on the viewpoint, and the user's stance on the viewpoint was always inserted after the human showed agreement to the premise. Accordingly, as indicated by the result of Q3, participants perceived the statements claimed in the proposed system to be more orderly than those in the flat condition.

To analyze whether the subjects' impressions of these points contributed to the persuasiveness rather than other subjects-side factors such as their characteristics, opinions,

Table 2 Results of multiple regression analysis with persuasiveness as the objective variable. VIF indicates the variance inflation factor. ΔR^2 indicates the drop of R^2 when this explanatory variable is not used.

Explanatory Variable	Standard Coefficient(β)	Standard Errors	t-value	VIF	ΔR^2
Intercept	0.000	0.0344	-	-	-
Q1: Dialogue Understanding	0.100	0.0397	2.525	1.353	0.00744
Q3: Orderly Statement	0.446	0.0491	9.074***	2.152	0.0797
Q4: Explicit Reason	0.222	0.0480	4.634***	1.983	0.0217
Q5: Reflectiveness	0.192	0.0435	4.416***	1.648	0.0184
KiSS-18	0.0386	0.0345	1.119	1.017	0.00146
Participant's Answer	0.0251	0.0345	0.727	1.017	0.000618

***: $p < .001$

or understanding of the discussion topics, multiple regression analysis was performed. In this analysis, the objective variable was the standardized score regarding "persuasiveness," while the explanatory variables were those regarding "orderly statement," "explicit reason," and "reflectiveness." Factors including the evaluation of a subject's characteristics in social aspects, as measured by the KiSS-18, the subject's attitude to the main proposition topic, and the self-evaluation of understanding of the given discussion dialogue were also included. Table 2 shows the result of the multiple regression analysis with the Bonferroni correction. We calculated a variance inflation factor (VIF) [32] to check the multicollinearity. If the VIF value of each explanatory variable is lower than ten, it is considered that multicollinearity did not occur [33]. Therefore, we consider that there is no problem of multicollinearity in this analysis. ΔR^2 indicates the drop of coefficient of determination R^2 calculated in the ablation test of the multiple regression. It was calculated using the equation $\Delta R^2 = R^2_{all} - R^2_{excluded}$ where R^2_{all} is the R^2 value calculated using all explanatory variables, while $R^2_{excluded}$ is also the value calculated using all explanatory variables excluding the one variable. Although a positive value in the ablation test indicated a potential contribution of the objective variable, according to the t-value, persuasiveness was significantly affected only by the following variables: orderly statement ($\beta = 0.446$, $p < .001$), explicit reason ($\beta = 0.222$, $p < .001$), and reflectiveness ($\beta = 0.192$, $p < .001$). Additionally, according to the standardized coefficient, the analysis suggested that the orderly statement variable was the most influential. The coefficient of determination R^2 was 0.562. Accordingly, as was predicted, the proposed system succeeded in generating dialogues that the subjects regarded as more persuasive in the hierarchical condition than in the flat one. This perceived improved persuasion was primarily impacted by the hierarchical dialogues' more orderly format, provision of reasons, and reflection of the user's utterances.

6. Conclusion

A method utilizing hierarchical argumentation structure was proposed to generate persuasive argumentative dialogues by choosing utterances based on a user's agreement and disagreement. The experimental results suggested that the dialogue generated by the proposed hierarchical method was more persuasive than the baseline flat structure method, which contrastingly did not alter its utterances according to the user's attitude. Limitations of the study include its reliance on an objective third person evaluation platform to evaluate persuasiveness as well as its use of solely supportive nodes which preclude rebuttals. To address these shortcomings, future works should aim to involve users' direct interaction with the system, wherein users' predilections regarding the argument topic are controlled. Additionally, non-supportive child nodes should be incorporated to permit rebuttals in the system's utterances. Such developments could further inform the ability for the proposed hierarchically structured system to inform programs for persuasive conversation strategies and computational argumentation generation structures.

Acknowledgments

This work was supported by JST ERATO Grant Number JPMJER1401, Japan.

References

[1] D. Schulman and T. Bickmore, "Persuading users through counseling dialogue with a conversational agent," Proceedings of the 4th International Conference on Persuasive Technology, pp.25:1–25:8, 2009.

[2] D. Walton, Argumentation Schemes, Cambridge University Press, 2008.

[3] J. Lawrence, F. Bex, C. Reed, and M. Snaith, "AIFdb: Infrastructure for the argument web," Computational Models of Argument (COMMA), pp.515–516, 2012.

[4] S. Modgil and J. McGinnis, "Towards characterising argumentation based dialogue in the argument interchange format," Argumentation in Multi-Agent Systems, vol.4946, pp.80–93, 2008.

[5] C. Reed, S. Wells, J. Devereux, and G. Rowe, "Aif+: Dialogue in the argument interchange format.," Proceedings of the 2nd international conference on Computational Models of Argument, pp.311–323, 2008.

[6] N. Rach, S. Langhammer, W. Minker, and S. Ultes, "Utilizing argument mining techniques for argumentative dialogue systems," Proceedings of the 9th international Workshop on spoken dialogue systems, vol.579, pp.131–142, 2018.

[7] K. Sakai, A. Inago, R. Higashinaka, Y. Yoshikawa, H. Ishiguro, and J. Tomita, "Creating large-scale of argumentation structures for dialogue systems," Proceedings of the 11th edition of the language resources and evaluation conference, pp.3975–3980, 2018.

[8] R. Higashinaka, K. Sakai, H. Sugiyama, H. Narimatsu, T. Arimoto, T. Fukutomi, K. Matsui, Y. Ijima, H. Ito, S. Araki, Y. Yoshikawa, H. Ishiguro, and Y. Matsuo, "Argumentative dialogue system based on argumentation structures," Proceedings of the 21st Workshop on the Semantics and Pragmatics of Dialogue, pp.154–155, 2017.

[9] K. Sakai, R. Higashinaka, Y. Yoshikawa, H. Ishiguro, and J. Tomita, "Introduction method for argumentative dialogue using paired question-answering interchange about personality," Proceedings of the 19th annual SIGDIAL Meeting on discourse and dialogue, pp.70–79, 2018.

[10] S.E. Toulmin, The uses of argument, Cambridge university press, 1958.

[11] C. Reed and G. Rowe, "Araucaria: software for argument analysis, diagramming and representation," International Journal on Artificial Intelligence Tools, vol.13, no.4, pp.961–980, 2004.

[12] D. Walton, Methods of argumentation, Cambridge University Press, 2013.

[13] A. Peldszus and M. Stede, "From argument diagrams to argumentation mining in texts: A survey," International Journal of Cognitive Informatics and Natural Intelligence, vol.7, no.1, pp.1–31, 2013.

[14] M. Lippi and P. Torroni, "Argumentation mining: State of the art and emerging trends," ACM Transactions on Internet Technology, vol.16, no.2, pp.10:1–10:25, 2016.

[15] M.-F. Moens, E. Boiy, R.M. Palau, and C. Reed, "Automatic detection of arguments in legal texts," Proceedings of the 11th international conference on Artificial intelligence and law, pp.225–230, 2007.

[16] B.K. Bal and P. Saint-Dizier, "Towards building annotated resources for analyzing opinions and argumentation in news editorials," Proceedings of the language resources and evaluation conference, pp.1152–1158, 2010.

[17] S. Rosenthal and K. McKeown, "Detecting opinionated claims in online discussions," Proceedings of the 6th IEEE International Conference on Semantic Computing, pp.30–37, 2012.

[18] K. Yanai, Y. Kobayashi, M. Sato, T. Yanase, T. Miyoshi, Y. Niwa, and H. Ikeda, "Debating artificial intelligence," Hitachi Review, vol.65, no.6, p.151, 2016.

[19] J. Lawrence, M. Snaith, B. Konat, K. Budzynska, and C. Reed, "Debating technology for dialogical argument: Sensemaking, engagement, and analytics," ACM Trans. Internet Technol., vol.17, no.3, pp.24:1–24:23, 2017.

[20] G. Rakshit, K.K. Bowden, L. Reed, A. Misra, and M. Walker, "Debbie, the debate bot of the future," Advanced Social Interaction with Agents - 8th International Workshop on Spoken Dialog Systems, IWSDS, vol.510, pp.45–52, 2017.

[21] D.T. Le, C.-T. Nguyen, and K.A. Nguyen, "Dave the debater: a retrieval-based and generative argumentative dialogue agent," Proceedings of the 5th Workshop on Argument Mining, pp.121–130, 2018.

[22] R. Marumoto, K. Tanaka, T. Takiguchi, and Y. Ariki, "Debate dialog for news question answering system 'nettv'—debate based on claim and reason estimation," International Workshop on Spoken Dialog Systems, IWSDS, vol.579, pp.389–396, 2018.

[23] F. Bex and D. Walton, "Combining explanation and argumentation in dialogue," Argument and Computation, vol.7, no.1, pp.55–68, 2016.

[24] A. Hunter, "Modelling the persuadee in asymmetric argumentation dialogues for persuasion," Proceedings of the 24th International Conference on Artificial Intelligence, pp.3055–3061, 2015.

[25] M. Thimm, "Strategic argumentation in multi-agent systems," KI - Künstliche Intelligenz, vol.28, no.3, pp.159–168, 2014.

[26] A. Rosenfeld and S. Kraus, "Strategical argumentative agent for human persuasion," Proceedings of the Twenty-second European Conference on Artificial Intelligence (ECAI), pp.320–328, 2016.

[27] K. Georgila and D. Traum, "Reinforcement learning of argumentation dialogue policies in negotiation," The 12th Annual Conference of the International Speech Communication Association (InterSpeech), pp.2073–2076, 2011.

[28] M. Koit, "Reasoning and communicative strategies in a model of argument-based negotiation," Journal of Information and Telecommunication, vol.2, no.3, pp.291–304, 2018.

[29] A. Kikuchi, Syakaiteki sukiru wo hakaru: KiSS-18 handbook [Measuring social skills: KiSS-18 handbook], Kawashima, Tokyo, 2007.

[30] T. Arimoto, Y. Yoshikawa, and H. Ishiguro, "Multiple-robot conversational patterns for concealing incoherent responses," International Journal of Social Robotics, vol.10, no.5, pp.583–593, 2018.

[31] C.R. Wolfe, M.A. Britt, and J.A. Butler, "Argumentation schema and the myside bias in written argumentation," Written Communication, vol.26, no.2, pp.183–209, 2009.

[32] D.W. Marquardt, "Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation," Technometrics, vol.12, no.3, pp.591–612, 1970.

[33] J.D. Curto and J.C. Pinto, "The corrected VIF (CVIF)," Journal of Applied Statistics, vol.38, no.7, pp.1499–1507, 2011.

Kazuki Sakai received his master's degree from Osaka University in 2017. He is a Ph.D. student in Osaka University. His research interests include human-robot interaction and dialogue system.

Ryuichiro Higashinaka received a B.A. in environmental information, a Master of Media and Governance, and a Ph.D. from Keio University, Kanagawa, in 1999, 2001, and 2008. He joined NTT in 2001. He is currently a senior distinguished researcher at NTT Media Intelligence Laboratories. His research interests include building question answering systems and spoken dialogue systems. From November 2004 to March 2006, he was a visiting researcher at the University of Sheffield in the UK. He received the Prize for Science and Technology of the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology in 2016. He is a member of the Japanese Society for Artificial Intelligence, the Information Processing Society of Japan, and the Association for Natural Language Processing.

Yuichiro Yoshikawa received the Ph.D. degree in engineering from Osaka University, Japan, in 2005. From 2005, he was a Researcher at Intelligent Robotics and Communication Laboratories, Advanced Telecommunications Research Institute International. From 2006, he has been a Researcher at Asada Synergistic Intelligence Project, ERATO, Japan Science and Technology Agency. From 2010, He has been an Associate Professor in the Graduate School of Engineering Science, Osaka University. From 2014, he has been a project coordinator of JST ERATO Ishiguro Symbiotic Human-Robot Interaction Project. He is a member of Japanese Society of Robotics, Japanese Society of Cognitive Science, the Virtual Reality Society of Japan, Japanese Society for Child and Adolescent Psychiatry, and Japanese Society of Pediatric Psychiatry and Neurology.

Hiroshi Ishiguro received a D. Eng. in systems engineering from the Osaka University, Japan in 1991. He is currently Professor of Department of Systems Innovation in the Graduate School of Engineering Science at Osaka University (2009-) and Distinguished Professor of Osaka University (2017-). He is also visiting Director (2014-) (group leader: 2002-2013) of Hiroshi Ishiguro Laboratories at the Advanced Telecommunications Research Institute and an ATR fellow. His research interests include sensor networks, interactive robotics, and android science.

Junji Tomita received the M.S. degree in computer science from Keio University, Kanagawa in 1997. He joined NTT in 1997. He was a visiting scholar at University of Washington in 2005. He worked for NTT Resonant Inc. from 2006 to 2017. He received the Ph.D. from Keio University in 2012. He is currently a senior research engineer at NTT Media Intelligence Laboratories. His research interests include natural language processing, information retrieval, and dialogue systems. He is a member of Information Processing Society of Japan, and a board member of the Data Base Society of Japan.