Utilizing Argument Mining Techniques for Argumentative Dialogue Systems

Niklas Rach, Saskia Langhammer, Wolfgang Minker, and Stefan Ultes

Abstract This work presents a pilot study for the application of argument mining techniques in the context of argumentative Dialogue Systems. We extract the argument structure of an online debate and show how it can be utilized to generate artificial persuasive dialogues in an agent-agent scenario. The interaction between the agents is formalized as argument game and the resulting artificial dialogues are evaluated in a user study by comparing them to human generated ones. The outcomes indicate that the artificial dialogues are logically consistent and thus show that the use of the employed argument annotation scheme was successful.

1 Introduction

Argumentation is an essential part of human conversation and is often employed in order to resolve conflicts or persuade an opponent. Enabling virtual agents to engage with humans (and each other) in a similar way (i.e. by exchanging arguments) is crucial for tasks such as reasoning, understanding new concepts and deducing new knowledge [8]. However, implementations of such argumentative systems have to overcome different barriers [19] and rely on knowledge about existing arguments and their formal relations to each other [2, 12, 14, 18]. With the large amount of arguments present online in different forms and from numerous sources, approaches to harness the same for the before mentioned systems are of particular interest. The field of argument mining [8, 7, 6] is concerned with the automatic extraction and analysis of argumentative structures from natural language sources and thus provides promising approaches for this task. In this work, we employ the annotation

Niklas Rach, Saskia Langhammer, Wolfgang Minker
Institute of Communication Engineering, Ulm University, Germany e-mail: niklas.rach@uni-ulm.de, wolfgang.minker@uni-ulm.de

Stefan Ultes
Department of Engineering, University of Cambridge, UK e-mail: su259@cam.ac.uk

scheme introduced in [15] to extract the argument structure of an online debate and show how it can be utilized in an argumentative Dialogue System.

We choose an agent-agent scenario as testbed and generate artificial dialogues from the resulting argumentation structure in order to analyze and evaluate our approach. Following the classification of [13], we focus on persuasive dialogues meaning that each agent has the goal to establish a convincing line of argumentation and to weaken the one of the opponent. To this end, the dialogue is formalized as argument game (see [10, 17] for an overview) based on the formal system presented in [9]. The evaluation is done by comparing the resulting artificial dialogues to human generated ones in a user study. The present work reports first steps in the implementation of the argumentative Dialogue System discussed in [11] and builds on the work presented in [4].

The remainder of the paper is as follows: Section 2 examines related work on argumentative Dialogue Systems. Section 3 introduces the employed argument mining scheme and discusses the textual source as well as the annotation. In Section 4, we introduce the architecture of the system in combination with the theoretical background of the respective components. Subsequently, in Section 5, the outcome and the evaluation of the corresponding user study are discussed. We close with a conclusion and a brief discussion of future work in Section 6.

2 Related Work

In this section we summarize related work on argumentative (Dialogue) Systems. Implementations of this kind are comparatively scarce due to several issues [19] that have to be solved or bypassed.

Two examples that are also based on argument games are presented in [2] and [18]. The first one implements a specification of the Toulmin Dialogue game, whereas the latter one is based on the Dialogue Model DE. Both systems assume a certain structure for the required database of arguments but in contrast to the present work, the focus lies not on its generation. Instead, exemplary argument structures are employed to illustrate the underlying principles. The Arvina system [5] allows users to exchange arguments present in the Argument Web [3] following the rules of a previously selected argument game. In addition, virtual agents are available to represent authors in the Argument Web by reflecting their arguments on a certain topic. Thus, these agents do not establish an individual line of argumentation, as in the herein discussed case.

Rosenfeld and Kraus [14] introduced a persuasive agent capable of learning an optimal strategy by means of Reinforcement Learning (RL) in a bipolar weighted argument framework. In contrast to the above mentioned systems, the interaction in this case is restricted to the exchange of available arguments, meaning that strategical moves (for instance questioning the validity of a previous argument) are not possible. The corpus of arguments is derived from human-human dialogues about the same topic by mapping utterances to arguments in the framework. Thus, an

inclusion of arguments from external sources as proposed in this work is not intended. Rakshit et al. [12] recently introduced an argumentative chat bot that relies on a corpus of annotated arguments generated from web resources. The system responses are derived by means of similarity measures between the user utterance and responses available in the corpus. Hence, each system response has to be included explicitly in the data. In contrast, the herein presented approach addresses the system response in the argument game framework, allowing the agents to respond to earlier utterances (and not just the latest) and to employ additional moves that have no corresponding instance in the data (for example challenging the validity of an argument).

3 Data and Annotation

In this section we discuss the annotation scheme and the textual source it is applied to. Our source of choice is a sample debate from the *Debatabase* of the idebate.org[1] website. The reasons for this choice are as follows: Firstly, idebate.org is operated by the International Debate Education Association (IDEA), a global network of organizations devoted to debating education. Hence, the debates offered here can be expected to meet certain quality standards regarding both form and content. Secondly, all debates presented here explore both sides of their respective topic. Lastly, all *Debatabase* debates adhere to a specific structure which both facilitates the quick screening for suitable candidates and potentially aids the argument annotation process later on. The sample debate employed in the scope of this work is concerned with the topic *Marriage is an outdated institution*. This choice is mostly due to the high amount of arguments provided by the *Debatabase* for this topic.

The employed annotation scheme was introduced by Stab et al. [15] for the analysis of written essays. It includes three argument components (*major claim, claim, premise*) and two directed relations (*support* and *attack*) between these components and is not tied to a specific domain. Thus, is provides every aspect required for the herein considered textual source. In addition, the resulting argument structure is compatible with the employed argument game (see Section 4) which makes the annotation scheme a reasonable choice for our task.

If a component ϕ_1 *supports* or *attacks* another component ϕ_2, we say that ϕ_2 is the target of ϕ_1 (or that ϕ_1 targets ϕ_2, respectively). A debate usually has one *major claim*, which formulates the overall topic around which the debate is built (here *Marriage is an outdated institution*) and is the only component that has no target. *Claims* are statements or assertions that express a certain opinion towards the *major claim* but require additional argumentative justification. Thus, *claims* can only target the *major claim*, not other *claims* or *premises*. An example *claim* from the herein examined debate is:

[1] https://idebate.org/debatabase (last accessed 09 January 2018)

Marraige does not provide any more of a stable environment for child rearing than a regular monogamous relationship.

A *premise* on the other hand provides reason for or against a *claim* or extends an already existing line of argumentation, meaning a *premise* can target the *major claim*, a *claim* or another *premise*. An example *premise* supporting the above mentioned *claim* is:

So many marriages end in divorce with the resulting splits affecting the children.

All argument components can target no more than one other component, but can be targeted by more than one and no argument component can target itself. This hierarchical structure of components allows for a representation of the annotated argumentative structure as a tree, where the argument components constitute the nodes and the argument relations constitute the edges. The arguments employed in the artificial dialogue can be constructed from this graph as a pair of nodes and their relation to each other. We denote the set of all arguments of this kind with *args*. The elements of *args* have the form $a = (\phi_1, \text{so } \phi_2)$ if ϕ_1 supports ϕ_2 and $b = (\phi_1, \text{so } \neg\phi_2)$ if ϕ_1 attacks ϕ_2. The construction of arguments from the graph is depicted in Figure 1. It is important to note that it is generally possible to build arguments

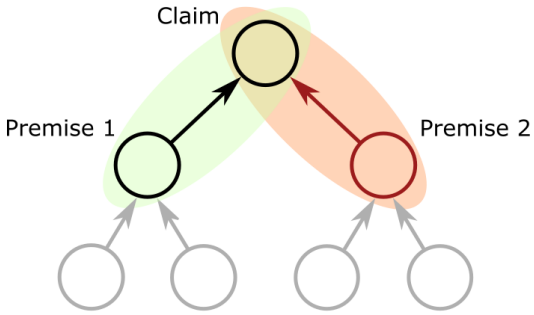

Fig. 1 Construction of arguments from the graph representing the annotated argument components. The green circle indicates an argument of the form *(premise 1, so claim)* the red circle a counterargument with the form *(premise 2,* so $\neg$*claim*)

with more than two components. Throughout this work we limit ourselves to the above described arguments consisting of two components, in order to ensure the best compatibility with the employed argument game framework.

The annotation was done by an expert based on the guidelines of [15] by first identifying argument components in the debate and secondly annotating the relations between them. The identification of overall topic and stance of the author included in the original work as a separate step are not required here, as both aspects are brought out by the structure of the debates. The annotation resulted in a total of 72 argument components (1 *major claim*, 10 *claims* and 61 *premises*) and their corresponding relations and was encoded in an OWL ontology [1] for further use.

In order to facilitate the Natural Language Generation of arguments in the artificial dialogues, the original annotated sentences were modified slightly to form a complete and reasonable utterance. Thus, implications are made explicit, references and citations were reformulated and expressions that are exclusively used in the debate format were adapted to the dialogue context (for example *"[...] as the opposition claims"* was changed to *"[...] as you claim"*).

4 The System Architecture

In the following, we describe the architecture of our system as well as the theoretical foundations it is based on. The core of the system consists of two agents (Alice and Bob) that argue about a certain topic. The interaction follows rules defined in the argument game [9] which determines the player to move and the available moves in each state of the dialogue. The set of available arguments for both agents is provided by the argument tree discussed in Section 3. In the current state of our work, each dialogue has a fixed length, i.e. a maximal number of turns after which the interaction terminates and thus, no termination criterion is employed, yet. After the game is finished, each move is subject to a template based Natural Language Generation (NLG) transforming the game moves into a natural language utterance. In the subsequent sections we discuss three aspects of the system in more detail.

4.1 The Argument Game

To structure the interaction between the agents, we utilize the formal system for computational persuasion of Prakken [9] that formalizes utterances in the dialogue as moves in a game. The possible moves for each agent have the form *claim*(ϕ), *argue*(a), *why*(ϕ), *retract*(ϕ), *concede*(ϕ) with ϕ an argument component in the graph discussed above and $a \in args$. It is worth noting that only two of these types introduce new content, i.e. new argument components to the dialogue (*claim* and *argue*), whereas the remaining three deal with components introduced earlier. In addition, the formalism introduces a protocol to determine the outcome of the game, the player to move and the list of available moves at each state. The outcome is tied to the termination criterion which is not employed here due to the finite length of the dialogues. Thus, we do not determine a winner (in the sense of the dialogue game) in the herein discussed scenario.

The list of available moves is determined by means of a relevance criterion for previous moves defining which of the latter ones can be addressed in the current state of the game. Thus, a response to earlier moves (and not just the latest one) is possible. We refer to responses of this kind as *topic switch*, since the focus is switched from the latest move to an earlier one. This allows to respond to a move more than once (if the required conditions are met) and gives the agents the possi-

bility to explore different branches of the argument tree in the same dialogue. Apart from the opening move (*claim*), each move either attacks a previous move (*argue*, *why*) or surrenders to a previous move (*concede*, *retract*).

For a detailed discussion of the framework we refer the interested reader to [9, 10]. It should be noted that the *claim* move is not to be confused with the *claim* component of the annotation scheme. In fact, the *claim* move is only employed to open the game and is thus always introduces the *major claim* component to the dialogue.

4.2 Agent strategy

The second issue is the agents strategy. As the framework restricts the agents only as much as necessary, a strategy to select the next move from the list of allowed ones is required as different moves lead to different outcomes. Throughout this work, we employ rules that lean to the argumentative agent profile described in [7]. The key assumption is, that the agent attacks whenever possible. Consequently, he only surrenders to an opponent move if there is no other option left. This choice is reasonable for the herein discussed case, as each participants goal is to convince the opponent and thus to strengthen the own and weaken the opponents line of argumentation whenever possible. To keep the dialogue focused on the current topic, we add a preference of moves that respond to the latest opponent move over a topic switch. In addition, we add a preference of *argue* over *why* moves in order to prevent an extensive use of the latter one. The rules thus read as:

- Attack if possible. If you do so,
  - If possible, attack the previous utterance of the opponent.
  - Prefer *argue* moves over *why* moves.
- If no attack is possible, *surrender*. If possible, *surrender* to the latest opponents move.

By using these rules, each agent identifies its next move from the set of possible moves. If there is more than one move fulfilling the same conditions, the next move is picked from this subset randomly.

4.3 Natural Language Generation

The NLG of the system relies on the original textual representation of the argument components. As discussed in Section 3, the annotated sentences were slightly modified to form a stand-alone utterance which serves as a template for the respective *argue* (and *claim*) move. In addition, a list of natural language representations for each additional type of move was defined. The explicit formulation is chosen from

this list randomly during the generation of the utterances. Moreover, topic switches are indicated by referencing formulations, followed by the textual representation of the referenced argument component as for example: *I'd like to go back to something you mentioned earlier. I think you said: [...]*. Again, the explicit formulation is picked randomly for each utterance.

Finally, we added additional connecting phrases like *That's not a valid point.* to some argue moves in order to increase the naturalness of the dialogue.

5 Outcome and Evaluation

In this section, we discuss the outcome of the overall system, i.e. the artificial dialogues between the two agents Alice and Bob, and evaluate the annotation in this context. To this end, we first examine one of the artificial dialogues as well as an excerpt of a human generated discussion and highlight structural differences. Subsequently, we discuss the findings of a user study comparing both types of dialogues. The focus herein lies on the logic consistency of the lines of argumentation established by the agents as this consistency depends directly on the annotated argument structure. From our perspective, this is a fair evaluation of both the annotation and application, as errors in one of the two or even both aspects would be reflected in the outcome.

5.1 Artificial Dialogue

In the following, we first show an exemplary artificial dialogue generated by the system by means of the annotated argument structure[2]. The utterances are generated by the NLG, whereas the content of the brackets indicates the respective move (*claim: c, argue: a, why: w*) including its index, followed by its target. For instance, (a2,a1) denotes argue move 2, addressing (the previous) argue move 1.

Alice: *"Marriage is an outdated institution."* (c1)

Bob: *"Why do you think that?"* (w1,c1)

Alice: *"The frequency and accessibility of divorce undermines the entire purpose of marriage."* (a1,w1)

Bob: *"Could you please elaborate?"* (w2,a1)

Alice: *"Marriage no longer leads to a stable or permanent relationship."* (a2,w2)

[2] Material reproduced from www.iedebate.org with the permission of the International Debating Education Association. Copyright ©2005 International Debate Education Association. All Rights Reserved

Bob: *"That's not a valid point. The purpose of marriage is not an eternal, unrelenting union, whether it is wanted or not."* (a3,a2)

Alice: *"Could you be more specific?"* (w3,a3)

Bob: *"The purpose of marriage is to foster a more stable relationship than would be possible without marital vows."* (a4,w3)

Alice: *"Could you please elaborate?"* (w4,a4)

Bob: *"Well, maybe we can come back to this point later. You said earlier: Marriage no longer leads to a stable or permanent relationship. I'm not sure I understand what you're getting at."* (w5,a2)

Alice: *"Our society no longer respects marriage as a permanent institution."* (a5,w5)

Bob: *"Could you be more specific?"* (w6,a5)

Alice: *"Serial monogamy is becoming ever more common."* (a6,w6)

Bob: *"Could you please elaborate?"* (w7,a6)

Alice: *"50% of all divorces in the UK go on to remarry according to the Office for National Statistics."* (a7,w7)

It should be noted that this dialogue does not include any surrendering move, as both agents prefer attack moves over surrender moves and the dialogue has fixed length, so no agent has run out of attack options yet. We recognize one topic switch (w5), that makes a new branch of the argument tree available for both agents and allows Alice to support one of her points (a2) with additional arguments. Moreover, a frequent use of *why* moves can be observed, due to the fact that the system up to now does not allow for chained arguments, i.e. an extended utterance covering multiple argument components. This is particularly evident in comparison with the excerpt of a human generated dialogue shown below[3], where the opposite can be observed:

Alice: *"Marriage is an outdated institution."*

Bob: *"Why do you think that?"*

Alice: *"According to the Office for National Statistics the average age, in the UK, to get married is approximately 30 years old. Life expectancy in the UK is approximately 80 years. This means the average marriage expects people to commit to maintain a certain way of life for a period that is longer than they have actually been alive. This goes hand in hand with the rise of social acceptability of people having more than one life partner in their life to show that either marriage is an unreasonable expectation of someone or a meaningless charade that is not actually expected to be maintained."*

[3] Material reproduced from www.iedebate.org with the permission of the International Debating Education Association. Copyright ©2005 International Debate Education Association. All Rights Reserved

Bob: *"This argument only works under the assumption that we live in a society where divorce does not exist. If a person enters into a marriage without full awareness of what they have committed to and later need to get out of that marriage, they are free to. Being able to leave a marriage, though, does not make marriage a meaningless charade, as you claim. It is still more difficult to leave a marriage than it is to leave a non-marital committed relationship and so it makes a big difference."*

We see that in this case, a single utterance encapsulates multiple argument components and that the dialogue as a whole consists of less turns. Moreover, only one *why* move is present in this case. A last difference in view of the line of argumentation is the fact, that in the human generated discussion some premises occur before the corresponding conclusion is present. Up to now this is not possible for virtual agents in our system as it could lead to inconsistent lines of argumentation and is only reasonable in combination with utterances containing multiple arguments. In order to enable the system to consistently argue this way, a modification of the employed argument game is required.

5.2 Survey

To validate the annotation and the herein discussed use of the resulting argument structure, we compared the two kinds of dialogues discussed above in a user study. Each participant was given one random instance out of the two possible categories (agent-agent or human dialogue). To include all aspects of the original debate, five human generated dialogues and 20 agent-agent dialogues were utilized as the argument density was higher in the human case. The 122 participants were from the UK and assigned randomly to one instance, resulting in a splitting of 54 participants rating the agent-agent case and 68 rating the human generated case. The rating was done on a five point scale from *completely disagree* (1) to *completely agree* (5) and 10 questions about the persuasiveness of the involved parties, logical consistency of the argumentation and an overall impression of the dialogue. The questions distinguish between the two agents Alice and Bob and ask about both of them separately:

- I was not convinced by Bob/Alice and how he/she presented his/her case. (Strat. Bob/Alice)
- It was always clear which previous utterance Bob/Alice addressed in his/her turn. (Prev. Bob/Alice)
- The arguments presented by Alice/Bob are logically consistent responses to the utterances they refer to. (Arg. Bob/Alice)
- Alice's/Bob's line of argumentation is not logically consistent. (Arg. line Bob/Alice)
- It was difficult to follow the line of argumentation throughout the debate. (Arg. line diff.)
- The whole debate is natural and intuitive. (Nat. and int.)

It should be noted that due to the different formulations the desired ranking is not always the highest. Table 1 shows the corresponding statistical results for all questions. Each line includes the median for the artificial dialogues (Agent), the human dialogues (Human) and the corresponding p value achieved with a Mann-Whitney-U test and all 120 ratings. As mentioned in the beginning, our focus in the context of this work lies on the questions assessing the logical consistency of the argumentation. These are in particular the questions asking for appropriateness of the arguments (Arg. Bob/Alice) and the questions that asses the complete line of argumentation (Arg. line Bob/Alice). The questions related to the agents strategy (Strat. Bob/Alice) were posed in order to decouple the rating of the dialogical behavior form the rating of the lines of argumentation. Thus, the corresponding results are not discussed further, here. We see that in the case of the overall consistency (Arg. line Bob/Alice) the ratings for the different scenarios are close to each other as the median is equal for both agents. Moreover, the case of Bob yields no significant difference whereas Alice is on the threshold of $p = 0.05$. For the single step rating (Arg. Bob/Alice), we see in both cases a significant difference between the human generated and the agent-agent dialogue. Nevertheless, the median is the same for the case of Bob and still above the neutral value of 3.0 for Alice. As mentioned earlier, we experienced a frequent use of *why* moves as well as some unintuitive changes of topic that may have lead to distraction and irritation of the participant. This is mostly reflected in the different ratings for the two last questions which asses the overall impression of the dialogue. As the argumentation was nevertheless in each case rated as consistent, we value the annotation scheme as an adequate approach to collect the data for systems of the herein discussed kind.

	Agent	Human	p
Strat. Bob	3.0	2.5	0.11
Strat. Alice	3.0	3.0	0.26
Prev. Bob	4.0	4.0	0.06
Prev. Alice	4.0	4.0	0.29
Arg. Bob	**4.0**	**4.0**	**0.02**
Arg. Alice	**3.5**	**4.0**	$\leq$ **0.01**
Arg. line Bob	**2.0**	**2.0**	**0.72**
Arg. line Alice	**2.0**	**2.0**	**0.05**
Arg. line diff.	3.0	2.0	≤ 0.01
Nat. and int.	2.0	4.0	≤ 0.01

Table 1 Median and p value for both agent-agent (Agent) and human-generated (Human) dialogues. Bold lines indicate questions related to the logic consistency of the argumentation.

6 Conclusion and Outlook

We have presented a Dialogue System application for argument mining techniques. We discussed the annotation of a written debate and how the resulting argument structure can be employed by virtual agents to play an argument game. Moreover, we have evaluated the resulting artificial dialogues in a user study. The results indicate that the consistency is prevailed in the artificial dialogues, although there is room for improvement in view of the naturalness of the same. We conclude that the use of the employed annotation scheme was successful but additional effort is required in order to enable a more natural and intuitive interaction.

Thus, future work will focus on multiple aspects. First of all, we aim for an extended database by including both additional topics and additional arguments for the present topic. The long term goal in this context is to automatically identify argument components and relations as investigated in [16] and thus utilize the full potential of argument mining in the herein discussed context. A second direction of interest is the dialogue management, i.e. the selection of arguments which will be optimized by means of Reinforcement Learning to provide a more intuitive and natural line of argumentation. The naturalness of the dialogue can in our opinion also be increased by a more advanced NLG. Finally an interaction of one of the agents with a human user is of interest, as this is eventually the goal of our system.

Acknowledgements This work is part of a project that is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 376696351.

References

1. Bechhofer, S.: Owl: Web ontology language. In: Encyclopedia of Database Systems, pp. 2008–2009. Springer (2009)
2. Bench-Capon, T.J.: Specification and implementation of toulmin dialogue game. In: Proceedings of JURIX, vol. 98, pp. 5–20 (1998)
3. Bex, F., Lawrence, J., Snaith, M., Reed, C.: Implementing the argument web. Communications of the ACM **56**(10), 66–73 (2013)
4. Langhammer, S.: A debating ontology for argumentative dialogue systems. Bachelor's thesis, Institute of Communication Engineering, Ulm University (2017)
5. Lawrence, J., Bex, F., Reed, C.: Dialogues on the argument web: Mixed initiative argumentation with arvina. In: COMMA, pp. 513–514 (2012)
6. Lippi, M., Torroni, P.: Argumentation mining: State of the art and emerging trends. ACM Transactions on Internet Technology (TOIT) **16**(2), 10 (2016)
7. Moens, M.F.: Argumentation mining: Where are we now, where do we want to be and how do we get there? In: Post-Proceedings of the 4th and 5th Workshops of the Forum for Information Retrieval Evaluation, p. 2. ACM (2013)
8. Palau, R.M., Moens, M.F.: Argumentation mining: the detection, classification and structure of arguments in text. In: Proceedings of the 12th international conference on artificial intelligence and law, pp. 98–107. ACM (2009)
9. Prakken, H.: On dialogue systems with speech acts, arguments, and counterarguments. In: JELIA, pp. 224–238. Springer (2000)

10. Prakken, H.: Formal systems for persuasion dialogue. The knowledge engineering review **21**(2), 163–188 (2006)
11. Rach, N., Minker, W., Ultes, S.: Towards an argumentative dialogue system (2017)
12. Rakshit, G., Bowden, K.K., Reed, L., Misra, A., Walker, M.: Debbie, the debate bot of the future. arXiv preprint arXiv:1709.03167 (2017)
13. Reed, C., Norman, T.: Argumentation machines: New frontiers in argument and computation, vol. 9. Springer Science & Business Media (2003)
14. Rosenfeld, A., Kraus, S.: Strategical argumentative agent for human persuasion. In: ECAI, pp. 320–328 (2016)
15. Stab, C., Gurevych, I.: Annotating argument components and relations in persuasive essays. In: COLING, pp. 1501–1510 (2014)
16. Stab, C., Gurevych, I.: Identifying argumentative discourse structures in persuasive essays. In: EMNLP, pp. 46–56 (2014)
17. Wells, S., Reed, C.A.: A domain specific language for describing diverse systems of dialogue. Journal of Applied Logic **10**(4), 309–329 (2012)
18. Yuan, T., Moore, D., Grierson, A.: A human-computer dialogue system for educational debate: A computational dialectics approach. International Journal of Artificial Intelligence in Education **18**(1), 3–26 (2008)
19. Yuan, T., Moore, D., Reed, C., Ravenscroft, A., Maudet, N.: Informal logic dialogue games in human–computer dialogue. The Knowledge Engineering Review **26**(2), 159–174 (2011)