Strategies in Dialogues: A Game-Theoretic Approach

Magdalena KACPRZAK [a], Marcin DZIUBIŃSKI [b], and Katarzyna BUDZYNSKA [c,d]
[a] *Faculty of Computer Science, Białystok University of Technology, Poland*
[b] *Institute of Informatics, University of Warsaw, Poland*
[c] *Institute of Philosophy and Sociology, Polish Academy of Sciences, Poland*
[d] *School of Computing, University of Dundee, UK*

Abstract. The aim of the paper is to propose a game-theoretic description of strategies available to players in dialogues. We show how existing dialogical systems can be formalized as Nash-style games, and how the game-theoretic concept of solutions (dominant strategies, Nash equilibrium) can be used to analyse these systems. Our first study, discussed in this article, describes the game DC introduced by Mackenzie.

Keywords. Formal dialogue systems, Strategies in dialogue games, Game theory, Nash equilibrium

1. Introduction

Formal dialogue systems explore agents' behaviour during the process of communication. The specification of these systems typically uses such elements as protocol, locution and commitment rules, which make it possible to investigate which moves and replies are legal in a particular type of dialogue game. Some of these systems examine dialogue types that meet certain postulates of rationality, such as the prohibition of fallacies (see e.g. [8,12,26]), while others concentrate on the requirements that dialogical interaction must meet in order for communication to serve specific goals, such as persuasion, negotiation, information-seeking, inquiry, and so on (see e.g. [25,16,21,22,9,1]).

This paper aims to demonstrate how this approach can be expanded using elements of game theory [15,14,24,6] so that we can investigate what interaction participants in a dialogue game should choose in order to behave not only *correctly* (i.e. according to the rules of the game), but also *successfully* (i.e. according to the goals and preferences of the agents). In other words, we will show how to apply a game-theoretic framework to study which interactions of a given protocol should be chosen so that the course of the dialogue will be most advantageous and rational for a given player.

Much attention is devoted in the literature to strategies in dialogue games [4,7,11,20]. Yuan et al. in [28] consider an adaptation of Moore's utilization of Mackenzie's game, DC. This study reveals several weaknesses of DC in preventing fallacious arguments and common errors. Black and Hunter's work [1] offers a dialogue-game-style protocol for each subtype of inquiry dialogue and provides a strategy that selects exactly one of the legal moves. Emele et al. [3] introduce a framework which allows players to build flexible and adaptive strategies for arguing and apply it to agents in information-seeking domains. The formal specification of strategies is discussed by Kacprzak and Budzynska in [10]. They propose a logical system $\mathscr{AG}_n$ which enables the representation and verification of strategies in dialogue games. Although much effort has been devoted to strategies in games, only a few works combine a Nash-style game-theoretic approach with dialogue games. Procaccia and Rosenschein [18] proposed a game-based argumentation framework, which combines the game-theoretic and argumentation-based styles of negotiation. Riveret et al. [23] continue this approach; they introduce and study their own game, focusing on subgame perfect equilibria. In [13] Matt and Toni proposed a game-theoretic approach to characterize argument strength. In [19] Rahwan and Larson define argumentation mechanism design and determine rules that ensure that players have no incentive to manipulate the outcome.

The contribution of this paper is the application of Nash-style game theory to specific, existing dialogical games in order to explore the nature of their dynamics. We demonstrate the method of such application using Mackenzie's DC system [12]. As noted by [27], this system is one of the most popular dialectical games, and there have been many implementations and variations of the original game (see e.g. [28]).

The application of the game-theoretic method consists of three steps. First, the selected dialogue system needs to be expressed in game-theoretic notions (see Sect. 2.1). It must then be supplemented with the specification of the system's properties, such as players' preferences or payoffs, which are necessary for exploiting game-theoretic tools (Sect. 2.2). We show that DC needs to be expanded to include the following elements: (1) a goal for the dialogue (in this paper we choose to explore two goals, pure persuasion and conflict resolution); (2) termination and outcome rules; and (3) players' preferences. The necessity of the second element is, however, not standard for this procedure but rather specific to DC, which, like some other early dialogue systems, did not specify the conditions under which a dialogue terminates or a player wins.

Finally, we study solutions for the DC system with pure persuasion and conflict resolution (Sect. 3). The concept of solution defines sets of strategy profiles which represent stable outcomes of the game. We investigate two types of solutions which determine the instructions for playing a given game: dominant strategies and Nash equilibrium. Given a selected specification of DC (e.g. player preferences), if there exists a solution in *dominant strategies*, then it provides the instructions of how to play a DC game in order to win regardless of how the opponent behaves during the dialogue. On the other hand, if such a solution does not exist, but there is a *Nash equilibrium*, then the player's victory is not guaranteed, but the Nash equilibrium provides instructions on how to play rationally. Imagine a game in which two players choose a red or black card. If both of them choose the same colour, then they both get \$1, and if they choose different colours, they get 0. Such a game has two Nash equilibria: red-red and black-black. This means that if a player i plays red, he has no guarantee of winning. Yet if he knows that the other player will keep choosing red, then it will be rational for i to begin to choose red as well.

To the best of our knowledge only a few works exploit Nash-style game theory for dialogue analysis, and even these few have a slightly different research focus from ours. The work of Rahwan and Larson [19] is closely related to our approach, but they focus on the mechanism design for constructing a dialogue game that satisfies the desired properties. In other words, they do not analyse the game-theoretic features, e.g. Nash equilibrium, of some existing game, but rather start from some assumed properties of a game and then aim to create a game that satisfies these properties. Similarly, the authors of [23,18] use games in an extensive form as we do, but study the game they construct. They consider other game-theoretic features as well, i.e. subgame perfect equilibria.

2. Translating DC into game-theoretic framework

2.1. The model

The model is a game in extensive form, formalizing the DC argumentation system created by Mackenzie [12]. When defining the model we use the following standard notation. Given a set Σ, the set of all finite sequences over Σ is denoted by Σ^* and the set of all infinite sequence over Σ is denoted by Σ^ω. The empty sequence is denoted by ε and the operation of concatenation is denoted by $\cdot$. Given sets A and B, $C \subseteq A$, $D \subseteq B$, and a function $f : A \to B$, we use $\overrightarrow{f}(C)$, to denote the image of C, and $\overrightarrow{f}^{-1}(D)$ to denote the inverse image of D. Before we define the game, we need to define the following parameters of the game: the set of statements, the set of locutions and the relation of immediate consequence on the set of statements.

Let S_0 be a non-empty and countable set called the *set of atomic statements*. The *set of statements*, $S[S_0]$, is a minimal set such that:

- $S_0 \subseteq S$.
- If $s \in S$, then $\neg s \in S$. (Negation)
- If $T \subseteq S$, then $\bigwedge T \in S$. (Conjunction)
- If $s, t \in S$, then $s \to t \in S$. (Conditional)

The *set of locutions*, $L[S_0]$, is then defined as follows:

$$L[S_0] = S[S_0] \cup \{\mathbb{Q}s : s \in S[S_0]\} \cup \{\mathbb{W}s : s \in S[S_0]\} \cup \{\mathbb{Y}s : s \in S[S_0]\} \cup \{\mathbb{R}s : s \in S[S_0]\},$$

where $\mathbb{Q}$ (question), $\mathbb{W}$ (withdrawal), $\mathbb{Y}$ (challenge), and $\mathbb{R}$ (resolution demand) are operators used for locution construction.

A relation of *immediate consequence*, $\mapsto \subseteq 2^{S[S_0]} \times 2^{S[S_0]}$ is a binary relation on the set of statements such that for all $s, t \in S$, $\{s, s \to t\} \mapsto \{t\}$.

Given a set of atomic statements, S_0, and a relation of immediate consequence, $\mapsto$, the *(argumentation) game* is a tuple $\Gamma_{[S_0, \mapsto]} = \langle P, \pi, H, T, (\precsim_i)_{i \in P}, (A_i)_{i \in P}, (\alpha_i)_{i \in P} \rangle$ where

- $P = \{\mathrm{W}, \mathrm{B}\}$ is the set of players.
- $H \subseteq L[S_0]^* \cup L[S_0]^\omega$ is the *set of histories*. A history is a (finite or infinite) sequence of locutions from $L[S_0]$. The set of finite histories in H is denoted by $\bar{H}$.
- $\pi : \bar{H} \to P \cup \{\varnothing\}$ is the *player function* assigning to each finite history the player who moves after it, or $\varnothing$, if no player is to move. The set of histories at which player $i \in P$ is to move is $H_i = \overrightarrow{\pi}^{-1}(i)$.
- $T = \overrightarrow{\pi}^{-1}(\varnothing) \cup (H \cap L[S_0]^\omega)$ is the set of *terminal histories*. A terminal history is a history after which no player is to move, hence it consists of the set of finite histories mapped to $\varnothing$ by the player function and the set of all infinite histories.
- $\precsim_i \subseteq T \times T$ is the *preference relation* of player i defined on the set of terminal histories.[1]

[1] The preference relation is a total preorder, i.e. it is total and transitive.

- $A_i = L[S_0]$ is the *set of actions* of player $i \in P$.
- $\alpha_i : H_i \to 2^{A_i}$ is the *admissible actions function* of player $i \in P$, determining the set of actions that i can choose from after history $h \in H_i$.

In what follows we will assume that the set of atomic statements S_0 is fixed and omit it, writing S rather than $S[S_0]$ and L rather than $L[S_0]$. We start by defining the properties of the player function. In the case of the DC system, $\pi(h) \in \{\mathrm{W}\varnothing\}$ if $|h|$ is even and $\pi(h) \in \{\mathrm{B}, \varnothing\}$ if $|h|$ is odd. Additionally, the rules of dialogue place restrictions on when the game can terminate: if $\pi(h \cdot l) = \varnothing$, then $l \notin \{\mathbb{Q}s, \mathbb{Y}s, \mathbb{R}s\}$. The remaining specification of game termination depends on the context in which the dialogue system is used, as discussed below.

Having defined the sets of actions for the players and the player function, we move on to define the admissible actions functions of the players. The function is determined by the rules of dialogue. In the case of the DC system, the rules of dialogue are defined using the notion of players' commitment sets. The *commitment set function* of player i is a function $C_i : L^* \to L$, assigning to each finite sequence of locutions $h \in L^*$ the *commitment set* $C_i(h)$ of i at h. The commitment set function of $i \in P$ is defined inductively at follows (where, given $i \in P$, $\{-i\} = P \setminus \{i\}$):[2]

CR_0 $\quad$ $C_{\mathrm{W}}(\varepsilon) = C_{\mathrm{B}}(\varepsilon) = \varnothing$.

$CR_{\mathbb{Q}}$ $\quad$ If $h \in L^*$, $s \in S$ and $i = \pi(h)$, then

$$C_i(h \cdot \mathbb{Q}s) = C_i(h)$$
$$C_{-i}(h \cdot \mathbb{Q}s) = C_{-i}(h).$$

$CR_{\mathbb{R}}$ $\quad$ If $h \in L^*$ and $s \in S$ and $i = \pi(h)$, then

$$C_i(h \cdot \mathbb{R}s) = C_i(h)$$
$$C_{-i}(h \cdot \mathbb{R}s) = C_{-i}(h).$$

$CR_{\mathbb{W}}$ $\quad$ If $h \in L^*$ and $s \in S$ and $i = \pi(h)$, then

$$C_i(h \cdot \mathbb{W}s) = C_i(h) \setminus \{s\}$$
$$C_{-i}(h \cdot \mathbb{W}s) = C_{-i}(h).$$

$CR_{\mathbb{Y}}$ $\quad$ If $h \in L^*$ and $s \in S$ and $i = \pi(h)$, then

$$C_i(h \cdot \mathbb{Y}s) = C_i(h) \cup \{\mathbb{Y}s\} \setminus \{s\}$$
$$C_{-i}(h \cdot \mathbb{Y}s) = C_{-i}(h) \cup \{s\}.$$

[2] In the definition we retain most of the labels of properties of $C_i(h)$ from [12]. The only difference is rules CR_S and $CR_{\mathbb{Y}S}$, which we have merged to $CR_{S+\mathbb{Y}S}$ for convenience of presentation.

$CR_{S+\mathbb{Y}S}$ If $h \in L^*$ and $s \in S$ and $i = \pi(h)$, then[3]

$$C_i(h \cdot s) = \begin{cases} C_i(h) \cup \{s, s \rightarrow s'\}, \text{ if } h = h' \cdot \mathbb{Y}s' \text{ for } h' \in L^*, s' \in S \\ C_i(h) \cup \{s\}, \qquad \text{otherwise.} \end{cases}$$

$$C_{-i}(h \cdot s) = \begin{cases} C_{-i}(h) \cup \{s, s \rightarrow s'\}, \text{ if } h = h' \cdot \mathbb{Y}s' \text{ for } h' \in L^*, s' \in S \\ C_{-i}(h) \cup \{s\}, \qquad \text{otherwise.} \end{cases}$$

The set of *immediate consequence conditionals* $ICC = \{\bigwedge T \rightarrow s : T \mapsto \{s\}\}$. A set of statements, $T \subseteq S$, is *immediately inconsistent* if there exists a finite subset $T' \subseteq T$ and a statement s such that $\neg s \in T'$ and $T' \mapsto \{s\}$. For more details see [12].

The admissible actions function α_i of player $i \in P$ is defined as follows. Given $h \in H_i$, $\alpha_i(h)$ is a maximal set of locutions satisfying the following:[4]

R_{Repstat} $\alpha_i(h) \cap C_1(h) \cap C_2(h) = \varnothing$.

R_{Imcon} If $\mathbb{W}s \in \alpha_i(h)$, then $s \notin ICC$.

R_{Quest} If $h = h' \cdot \mathbb{Q}s$, then $\alpha_i(h) = \{s, \neg s, \mathbb{W}s\}$.

R_{LogChall} If $\mathbb{Y}s \in \alpha_i(h)$, then $s \notin ICC$.

R_{Chall} If $h = h' \cdot \mathbb{Y}s$, then $\alpha_i(h) =$
$\{\mathbb{W}s\} \cup \{s' \in S : \mathbb{Y}s' \notin C_{-i}(h' \cdot \mathbb{Y}s)\} \cup$
$\{\mathbb{R}(\bigwedge T \rightarrow s) : s \in S, T \subseteq C_{-i}(h' \cdot \mathbb{Y}s) \text{ and } T \mapsto \{s\}\}$.

R_{Resolve} If $\mathbb{R}s \in \alpha_i(h)$, then either

- $s = \bigwedge T$, T is immediately inconsistent, and $T \subseteq C_{-i}(h)$.
- $s = \bigwedge T \rightarrow u$, $s \in ICC$, $T \subseteq C_{-i}(h)$, and either $h = h' \cdot \mathbb{W}u$ or $h = h' \cdot \mathbb{Y}u$.

$R_{\text{Resolution}_{\bigwedge}}$ If $h = h' \cdot \mathbb{R}(\bigwedge T)$, then $\alpha_i(h) = \{\mathbb{W}s : s \in T\}$.

$R_{\text{Resolution}_{\bigwedge\rightarrow}}$ If $h = h' \cdot \mathbb{R}(\bigwedge T \rightarrow s)$, then $\alpha_i(h) = \{\mathbb{W}s' : s' \in T\} \cup \{s\}$.

The set of histories, H, is the maximal set of sequences from $L^* \cup L^\omega$ satisfying the following:

- $\varepsilon \in H$.
- For any $h_1 \cdot h_2 \in H$ with $h_1 \in L^*$ and $h_2 \in L^* \cup L^\omega$, $h_1 \in H$.
- For any $h_1 \cdot s \cdot h_2 \in H$ with $h_1 \in L^*$, $h_2 \in L^* \cup L^\omega$ and $s \in L$, $s \in \alpha_{\pi(h_1)}(h_1)$.

Two elements of the game are left undefined. These are player preferences and the termination rules that describe which finite histories are mapped to $\varnothing$.

Note that every strategy profile $\bar{S} = (S_{\text{W}}, S_{\text{B}})$ determines a unique terminal history $h_{\bar{S}}$ such that for each strategy $s \in \mathbf{S}_{\text{W}} \cup \mathbf{S}_{\text{B}}$, finite history $h' \in \bar{H}$ and history $h'' \in H$ with $h_{\bar{S}} = h' \cdot a \cdot h''$, $a = S_{\pi(h')}(h')$. Player $i \in P$ prefers strategy profile $\bar{S}$ to strategy profile $\bar{S}'$, $\bar{S}' \precsim_i \bar{S}$, if $h_{\bar{S}'} \precsim_i h_{\bar{S}}$. Thus to define players' preferences we need to define their preferences on terminal histories.

The definition of termination rules and preferences on terminal histories are independent of the dialogue system. Rather, they depend on the type of dialogue the system

[3] The statement $h = h' \cdot l$ (where l is a locution) states that: (1) h is a non-empty sequence (of locutions), (2) the last locution in h is l, and (3) h' is a prefix of h shorter by (the last) one locution.

[4] Again, we retain the original labels of the corresponding rules of dialogue given in [12].

is applied to and perhaps some other application-dependent considerations (e.g. it could be that players prefer histories which are shorter, as long as they attain their objectives in the end).

In the following subsection we use an example of a persuasion dialogue to illustrate how these two components can be defined, thus completing the definition of the game.

2.2. Persuasion

Persuasion dialogues are dialogues aimed at resolving conflicts of opinion between at least two participants [25]. In what follows we restrict our attention to two participants only. We also consider particular types of persuasion dialogue called *pure persuasion* and *conflict resolution*.

A conflict of opinion is with regard to a statement, $t \in S$, called a *topic*. One of the players plays the role of the *proponent* of t, holding a positive view on t. The other player plays the role of *opponent*, doubting t. According to [25], the conflict is resolved if all parties share the same point of view. If the dialogue ends, either one of the players is the winner of the dialogue (in which case the other player is the loser), or the dialogue ends in a tie, in which case neither of the players is a winner or a loser. If the dialogue does not end, then neither of the players is a winner or a loser. The proponent wins if both players hold the topic in their commitment sets (in which case the opponent loses). The opponent wins if neither player holds the topic in his commitment set. A formal definition of persuasion dialogues, including the definition of the game outcomes, can be found in [17].

By combining the DC system with pure persuasion we can complete the definition of the argumentation game Γ as follows. The first mover, player W, is the proponent and the second mover, player B, is the opponent. According to the description above, the game should end if one of the following happens: (i) The commitment sets of both players contain the topic t, or (ii) neither player's commitment set contains the topic t. Since according to the rule $CR_{S+\mathbb{Y}_S}$ the most recent statements of any player are added to the commitment sets of both players, we adjust these rules of termination by allowing the next mover to respond and withdraw these statements from his set of commitments. If he is unable to do so, the game will terminate after his move. Similarly, if at any point in the game neither player's commitment set contains t, the next mover has the opportunity to add t to at least one of the commitment sets. If he is unable to do so, the game terminates after his move. Formally, this amounts to a definition of finite terminal histories, which are defined as follows. A finite history $h \in \bar{H}$ is terminal, i.e. $\pi(h) = \varnothing$, if one of the following conditions is satisfied:

$$T_{\text{prop}} : |h| \text{ is even and } t \in C_{\text{W}}(h) \cap C_{\text{B}}(h), \quad T_{\text{opp}} : |h| \text{ is odd and } t \notin C_{\text{W}}(h) \cup C_{\text{B}}(h).$$

Note that in the DC system the set of admissible actions at each non-terminal history is non-empty. Therefore a finite history can be terminal only if one of the above conditions is satisfied.

Having defined finite terminal histories we will now define the preferences of the players. Let $H_{\text{W}}^{\text{win}}$ denote the set of finite histories for which condition T_{prop} is satisfied and let $H_{\text{B}}^{\text{win}}$ denote the set of finite histories for which condition T_{opp} is satisfied. Set $H_{\text{W}}^{\text{win}}$ contains the terminal histories at which player W, the proponent, is the winner, and set $H_{\text{B}}^{\text{win}}$ contains the terminal histories at which player W, the opponent, is the winner.

We assume the following preference relation on terminal histories. Given $h_1, h_2 \in T$,

$$h_1 \precsim_{\mathrm{W}} h_2 \text{ if } h_2 \in H_{\mathrm{W}}^{\mathrm{win}} \text{ or } h_2 \notin H_{\mathrm{B}}^{\mathrm{win}} \text{ and } h_1 \in H_{\mathrm{B}}^{\mathrm{win}}, \text{ and}$$
$$h_1 \precsim_{\mathrm{B}} h_2 \text{ if } h_2 \in H_{\mathrm{B}}^{\mathrm{win}} \text{ or } h_2 \notin H_{\mathrm{W}}^{\mathrm{win}} \text{ and } h_1 \in H_{\mathrm{W}}^{\mathrm{win}}.$$

In other words, each player prefers a terminal history at which he wins to that at which he does not win, and each player prefers a history at which the opponent does not win to one at which the opponent wins.

Note that the above-defined preferences of the players and termination rules imply that a rational proponent (player W) should always begin with an action that results in the topic t being added to the commitment set of at least one of the players.

Another type of persuasive dialogue is a conflict resolution dialogue, in which both players attempt to reach an outcome whereby either both of them have the topic t in their commitment sets or both have the negation of the topic, $\neg t$, in their commitment sets [2]. Conflict resolution dialogues begin with a *conflict of opinion*, i.e. the proponent holds the topic in his commitment set while opponent holds the negation of the topic in his commitment set. To model this situation we need to adjust the rules of the DC system so that the initial commitment sets are not empty. Thus we drop rule CR_0 and assume instead that $C_{\mathrm{W}}(\varepsilon) = \{t\}$ and $C_{\mathrm{B}}(\varepsilon) = \{\neg t\}$.

The game should end if one of the following happens: (i) The commitment sets of both players contain t, or (ii) the commitment sets of both players contain $\neg t$. As in the case of pure persuasion, we adjust the rules of termination by allowing the next mover to respond. Finite terminal histories are defined as follows. A finite history $h \in \bar{H}$ is terminal, i.e. $\pi(h) = \varnothing$, if one of the following conditions is satisfied:

$$T_{\mathrm{prop}} : |h| \text{ is even and } t \in C_{\mathrm{W}}(h) \cap C_{\mathrm{B}}(h), \quad T_{\mathrm{opp}} : |h| \text{ is odd and } \neg t \in C_{\mathrm{W}}(h) \cap C_{\mathrm{B}}(h).$$

As noted above, a finite history can be terminal only if one of the above conditions is satisfied.

Having defined finite terminal histories we move on to defining the preferences of the players. Let $H_{\mathrm{W}}^{\mathrm{win}}$ denote the set of finite histories for which condition T_{prop} is satisfied and let $H_{\mathrm{B}}^{\mathrm{win}}$ denote the set of finite histories for which condition T_{opp} is satisfied. We assume the following preference relation on terminal histories. Given $h_1, h_2 \in T$,

$$h_1 \precsim_{\mathrm{W}} h_2 \text{ if } h_2 \in H_{\mathrm{W}}^{\mathrm{win}} \text{ or } h_2 \in H_{\mathrm{B}}^{\mathrm{win}} \text{ and } h_1 \notin H_{\mathrm{W}}^{\mathrm{win}}, \text{ and}$$
$$h_1 \precsim_{\mathrm{W}} h_2 \text{ if } h_2 \in H_{\mathrm{W}}^{\mathrm{win}} \text{ or } h_2 \in H_{\mathrm{B}}^{\mathrm{win}} \text{ and } h_1 \notin H_{\mathrm{W}}^{\mathrm{win}}.$$

In other words, each player prefers a terminal history at which he wins to any other history, and each player prefers a history at which the opponent wins to one at which there is no winner.

3. Strategies and solutions

3.1. Definitions

A *strategy* of a player i is a function from player i's histories to the set of actions $S_i : H_i \to L$, such that for all $h \in H_i$, $S_i(h) \in \alpha_i(h)$. Thus a strategy is a contingent plan that determines a player's move at each of his histories. The set of strategies of player i is denoted by $\mathbf{S}_i$. A *strategy profile* $\bar{S} = (S_i, S_{-i})$ is a pair of strategies chosen by each of the players, $\bar{S} \in \mathbf{S}_i \times \mathbf{S}_{-i}$. Solution concepts define sets of strategy profiles which represent stable outcomes of the game. Below we define two basic solution concepts and illustrate them in the context of pure persuasion and conflict resolution.[5]

A strategy S_i is *dominant* for player i if for all $S_i' \in \mathbf{S}_i$ and all $S_j \in \mathbf{S}_j$ with $j = -i$,

$$(S_i', S_j) \precsim_i (S_i, S_j).$$

A strategy is *strictly dominant* if the property above holds with strict inequality. A strategy profile $\bar{S} = (S_i, S_{-i})$ is a *solution in (strictly) dominant strategies* iff S_i is dominant for player i and S_{-i} is dominant for player $-i$.[6]

A strategy profile $\bar{S} = (S_i, S_{-i})$ is a *Nash equilibrium* if for all $i \in P$ and for all $S_i' \in \mathbf{S}_i$,

$$(S_i', S_{-i}) \precsim_i (S_i, S_{-i}).$$

Note that if the game has a solution in dominant strategies, then it also has a Nash equilibrium (but the reverse is not necessarily true).

3.2. Solutions to pure persuasive dialogue with the DC system

Consider a game $\Gamma_{[S_0, \mapsto]}$ defined for some set of atomic statements S_0 and a relation of immediate consequence $\mapsto$. Suppose that players' preferences and finite terminal histories are defined as for a persuasive dialogue.

Suppose first that the topic, t, is an immediate consequence conditional, $t \in ICC$. Consider a strategy $s_1 \in S_{\mathrm{W}}$ such that $s_1(\varepsilon) = t$. Note that after history $h = t$ the commitment sets of both players contain t and, since $t \in ICC$, the opponent cannot remove t from the commitment set of either of the players. Therefore the game ends in two rounds and the outcome most preferred by player W is obtained. Thus any s_1 as described above is a dominant strategy of player W, the proponent. Moreover, for any $s_2 \in S_{\mathrm{B}}$, the outcome of the game from strategy profile $\bar{s} = (s_1, s_2)$ is the same: the commitment sets of both players contain t. Thus no player can obtain a strictly preferred outcome by changing his strategy. Hence $\bar{s}$ is a Nash equilibrium of the game and leads to an outcome whereby the proponent wins.

Let us next suppose that the topic, t is not an immediate consequence conditional, $t \notin ICC$. Let $s_2 \in S_{\mathrm{B}}$ be a strategy such that for any non-terminal history h with $t \in C_{\mathrm{W}}(h) \cap C_{\mathrm{B}}(h)$ and $|h|$ odd, $s_2(h) = \mathbb{W}t$. Note that in this case $\mathbb{W}t \in \alpha_{\mathrm{B}}(h)$, because the last action of h must have been t. On the other hand, let $s_1 \in S_{\mathrm{W}}$ be a strategy such that for any non-terminal history h with $t \notin C_{\mathrm{W}}(h) \cup C_{\mathrm{B}}(h)$, $s_1(h) = t$. Note that the strategy profile $\bar{s} = (s_1, s_2)$ is a Nash equilibrium of the game. This is because none of the players can obtain his most preferred outcome by changing his strategy. Note that this equilibrium results in an infinite history.

3.3. Solutions to conflict resolution (persuasive) dialogue with the DC system

Consider a game $\Gamma_{[S_0, \mapsto]}$ defined for some set of atomic statements, S_0, and a relation of immediate consequence, $\mapsto$. Suppose that players' preferences and finite terminal histories are defined as for a persuasive dialogue.

[5] Other solution concepts, omitted here for readability, include subgame-perfect equilibria, equilibria with the Markov property (i.e. dependent on the last move by the opponent only), and equilibria where the players' choices at every history depend on the commitment set rather than the entire history.

[6] Note that the solution in strictly dominant strategies, if it exists, is unique.

We will consider three cases separately: $t \in ICC$, $\neg t \in ICC$ and $\{t, \neg t\} \cap ICC = \varnothing$. In the first case the proponent has a dominant strategy s_1, and no matter what the opponent does, when the proponent uses s_1 the game ends in two rounds (as in the case of persuasive dialogue). The situation is similar in the case where $\neg t \in ICC$. Note, however, that as the second mover, player B may not be able to state $\neg t$ until it becomes an admissible action. Thus a dominant strategy of player B could be a strategy s_2 such that at every history h at which $t \in C_W(h) \cap C_B(h)$, $s_2(h) = \mathbb{W}t$, at every other history h at which stating $\neg t$ is admissible, $s_2(h) = \neg t$, and at every other history h, $s_2(h) \in \alpha_B(h)$. The proponent's best response to strategy s_2 is any strategy s_1 which allows the opponent to state t at some stage, since any other response to s_2 would result in an infinite terminal history, which is preferred less by W than a finite terminal history with $\neg t \in C_W(h) \cap C_B(h)$. Thus (s_1, s_2) with s_1 being the best response to s_2 is a Nash equilibrium.

Finally, suppose that neither t nor $\neg t$ is an immediate consequence conditional. In this case there are two types of Nash equilibria: one in which player W's most preferred outcome is attained and another in which player B's most preferred outcome is attained. For example, any strategy profile (s_1, s_2) as described above for the case where $\neg t \in ICC$ is a Nash equilibrium of the game and attains player B's most preferred outcome. Similarly, there is an equilibrium where player W's most preferred outcome is attained. The situation in this game is reminiscent of the Battle of the Sexes [19], in which both players need to coordinate on one of the two outcomes and the payoffs from each of these outcomes are non-symmetric.

3.4. Illustration

In order to illustrate the results presented in the paper, let us consider two games. The first is a game $\Gamma_{[S_0, \mapsto]}$ such that $\{p, q, r\} \subseteq S_0$ and $\{p \rightarrow q, q \rightarrow p\} \subseteq \mapsto$. We assume that this game is a pure persuasion game, for which terminal histories and the preference relation defined on these histories were given in Section 2.2.

Consider two versions of this game. In version A, $p \rightarrow q$ is a topic which belongs to ICC. Player W is a proponent of this topic and has a dominant strategy s_1 such that $s_1(\varepsilon) = (p \rightarrow q)$. Observe that the outcome of the game for strategy profile $\bar{s} = (s_1, s_2)$ for any strategy s_2 of player B is the same. Therefore, this strategy profile is a Nash equilibrium. Table 1 shows the play in line with this strategy. In this play, W starts with statement $p \rightarrow q$. In move 2, player B has no legal move removing $p \rightarrow q$ from the set of his commitments and loses.

move number	locution of W	locution of B	C_W	C_B
1	$p \rightarrow q$	-	$p \rightarrow q$	$p \rightarrow q$
2	-	no move	$p \rightarrow q$	$p \rightarrow q$

Table 1. A play for pure persuasion - version A

In the version B of the game, the proponent W defends a topic r which is not in ICC. Players W and B have strategies s_1 and s_2, respectively, such that (a) for any non-terminal history h with $r \notin C_W(h) \cup C_B(h)$, $s_1(h) = r$, and (b) for any non-terminal history h with $r \in C_W(h) \cap C_B(h)$ and $|h|$ odd, $s_2(h) = \mathbb{W}r$. In other words, the proponent states r whenever r is not a common commitment of both players and the opponent withdraws r from his commitment set whenever it appears there. The play given in Table 2 shows how the opponent can postpone his defeat. In this case, the strategy profile $\overline{s} = (s_1, s_2)$ also determines a Nash equilibrium.

move number	locution of W	locution of B	C_{W}	C_{B}
1	r	-	r	r
2	-	$\mathbb{W}r$	r	-
3	r	-	r	r
4	-	$\mathbb{W}r$	r	-
...	...	...	...	...

Table 2. A play for pure persuasion - version B

The second game we consider is a version of the Battle of the Sexes game (cf. [19]) in which the couple – Wilma (W) and Brian (B) – wants to decide on their plans for the day. Brian thinks they should go to a soccer match, while Wilma thinks they should attend the ballet. Wilma prefers the ballet to the soccer, but would still rather go to a soccer match than stay at home. Similarly, Brian prefers the soccer match to the ballet, but prefers the ballet to staying home.

In this game, we assume the set of atomic statements to be such that $\{p,q,r,t\} \subseteq S_0$, where p states that they should go to the ballet, q states that they should go to the soccer match, r states that Wilma is too sick for the outdoors, and t states that Brian's ex-wife will be at the ballet. Moreover, $\{q \rightarrow \neg p, p \rightarrow \neg q, r \rightarrow \neg q, t \rightarrow \neg p\} \subseteq \mapsto$. This game is a conflict resolution game with the topic p. Wilma is the proponent of p and Brian is the opponent of p. The terminal histories and preference relation on terminal histories for a conflict resolution dialogue were defined in Section 2.2.

First consider the following strategy profile $\overline{s} = (s_1, s_2)$, where s_1 is a strategy of Wilma's in which she states p and r as soon as it is possible and s_2 is a strategy of Brian's in which he states t, attacking Wilma's statement r. More formally, s_1 is a strategy such that (a) $s_1(\varepsilon) = p$ and (b) for any non-terminal history h, if $r \in \alpha_{\mathrm{W}}(h)$ then $s_1(h) = r$, while s_2 is a strategy such that for any non-terminal history h, if $|h|$ is odd, $h = h' \cdot r$, and $t \in \alpha_{\mathrm{B}}(h)$ then $s_2(h) = t$. A play in line with this strategy profile is given in Table 3.

move number	locution of W	locution of B	C_{W}	C_{B}
1	p	-	p	p
2	-	$\mathbb{W}p$	p	-
3	r	-	p,r	r
4	-	t	p,r,t	r,t
5	$t \rightarrow \neg p$	-	$p,r,t,t \rightarrow \neg p$	$r,t,t \rightarrow \neg p$
6	-	$\mathbb{R}(p \wedge t \wedge (t \rightarrow \neg p))$	$p,r,t,t \rightarrow \neg p$	$r,t,t \rightarrow \neg p$
7	$\mathbb{W}p$	-	$r,t,t \rightarrow \neg p$	$r,t,t \rightarrow \neg p$

Table 3. A play for conflict resolution - version A

The play starts with Wilma's statement that they should go to the ballet. Brian withdraws this statement from his commitment set. Next, Wilma states that she is too sick for the outdoors. Brian replies that his ex-wife will be at the ballet. Then Wilma states that this fact rules out going to the ballet and Brian asks for a resolution of the conflict.

Finally, Wilma withdraws her statement that they should go to the ballet. The dialogue finishes in a state in which topic p is neither in the commitment set of player W nor in that of player B. This outcome is not the most preferred by the players. Note that if Wilma states that she is too sick, Brian should not advance the argument against the ballet. Therefore, the strategy profile $\overline{s}$ is not a Nash equilibrium.

Now, consider the strategy profile $\overline{s}' = (s'_1, s'_2)$ where s'_1 is a strategy of Wilma's in which she states p and r as soon as it is possible and s'_2 is a strategy of Brian's in which he states q as soon as it is possible, never attacking Wilma's statement by r stating t. More formally, s_1 is a strategy such that (a) $s_1(\varepsilon) = p$ and (b) for any non-terminal history h, if $r \in \alpha_{\mathrm{W}}(h)$ then $s_1(h) = r$, while s_2 is a strategy such that for any non-terminal history h, (a) if $q \in \alpha_{\mathrm{B}}(h)$ then $s_2(h) = q$, and (b) if $h = h' \cdot r \cdot h''$ then $s_2(h) \neq t$. A play in line with this strategy profile is given in Table 4.

move number	locution of W	locution of B	C_{W}	C_{B}
1	p	-	p	p
2	-	$\mathbb{W}p$	p	-
3	r	-	p,r	r
4	-	q	p,q,r	q,r
5	$r \to \neg q$	-	$p,q,r,r \to \neg q$	$q,r,r \to \neg q$
6	-	$\mathbb{R}(q \wedge r \wedge (r \to \neg q))$	$p,q,r,r \to \neg q$	$q,r,r \to \neg q$
7	$\mathbb{W}q$	-	$p,r,r \to \neg q$	$q,r,r \to \neg q$
8	-	$\mathbb{W}q$	$p,r,r \to \neg q$	$r,r \to \neg q$
9	p	-	$p,r,r \to \neg q$	$p,r,r \to \neg q$
10	-	no move	$p,r,r \to \neg q$	$p,r,r \to \neg q$

Table 4. A play for conflict resolution - version B

This play begins with Wilma's statement that they should go to the ballet. Brian withdraws this statement from his commitment set. Next, Wilma states that she is too sick for the outdoors. Brian replies that they should go to the soccer match. Then Wilma states that they cannot go to the soccer match since she is too sick. So, Brian asks for a resolution of the conflict. Wilma then withdraws her statement that they should go to the soccer match. Brian, not wanting to put his wife at risk, also gives up the idea of going to the soccer match. Then Wilma repeats her statement that they should go to the ballet and finally Brian agrees, i.e. he makes no move removing p from his commitment set. The strategy profile $\overline{s}' = (s'_1, s'_2)$ described above is a Nash equilibrium of the game (i.e. no player wants to change his/her strategy, assuming that the other player does not) and it achieves Wilma's most preferred outcome.

Conclusions and Future Work

In this paper we show how existing dialogue systems can be formalized in terms of game theory. As an example we take Mackenzie's DC argumentation system and represent it as a game in extensive form. Furthermore, we show how solutions in dominant strategies and solutions in a Nash equilibrium can be used for the analysis of dialogue systems. In our future work, we plan to examine other types of dialogue systems, including systems containing an abstract argumentation framework [5], as well as dialogue games with incomplete information.

References

[1] E. Black and A. Hunter. An inquiry dialogue system. *Autonomous Agents and Multi-Agent Systems*, 19(2):173–209, 2009.

[2] K. Budzynska and K. Debowska. Dialogues with conflict resolution: goals and effects. In *Aspects of Semantics and Pragmatics of Dialogue. SemDial 2010*, pages 59–66, 2010.

[3] D. E. Chukwuemeka, F. Guerin, T. J. Norman, and P. Edwards. A framework for learning argumentation strategies. In *Proceedings of the Third International Workshop on Argumentation in Multi-Agent Systems*, pages 151–154, 2006.

[4] J. Devereux and C. Reed. Strategic argumentation in rigorous persuasion dialogue. In P. McBurney, I. Rahwan, S. Parsons, and N. Maudet, editors, *ArgMAS*, volume 6057 of *Lecture Notes in Computer Science*, pages 94–113. Springer, 2009.

[5] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and *n*-person games. *Artificial Intelligence*, (77):321–357, 1995.

[6] M. Dziubiński and S. Goyal. Network design and defence. *Games and Economic Beh.*, 79(1):30–43, 2013.

[7] S. Gabrielli, R. Maimone, P. Forbes, and S. Wells. Exploring change strategies for sustainable urban mobility. *CHI 2013 Designing Social Media for Change Workshop*, 2013.

[8] C. L. Hamblin. *Fallacies*. Methuen and Co. Ltd, 1970.

[9] A. Hussain and F. Toni. Bilateral agent negotiation with information-seeking. In *Proc. of the 5th European Workshop on Multi-Agent Systems*, 2007.

[10] M. Kacprzak and K. Budzynska. Reasoning about dialogical strategies. In M. Graa, C. Toro, R. J. Howlett, and L. C. Jain, editors, *KES (Selected Papers)*, volume 7828 of *Lecture Notes in Computer Science*, pages 171–184. Springer, 2012.

[11] K. Larson and I. Rahwan. Welfare properties of argumentation-based semantics. In *Proceedings of the 2nd COMSOC*, 2008.

[12] J. D. Mackenzie. Question-begging in non-cumulative systems. *J. of Phil. Logic*, 8:117–133, 1979.

[13] P.-A. Matt and F. Toni. A game-theoretic measure of argument strength for abstract argumentation. In S. Hlldobler, C. Lutz, and H. Wansing, editors, *JELIA*, volume 5293 of *Lecture Notes in Computer Science*, pages 285–297. Springer, 2008.

[14] J. F. Nash, Jr. Equilibrium Points in *n*-Person Games.

[15] J. V. Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[16] H. Prakken. Formal systems for persuasion dialogue. *The Knowledge Eng. Review*, 21:163–188, 2006.

[17] H. Prakken. Models of persuasion dialogue. In *Argumentation in AI*, pages 281–300. Springer, 2009.

[18] A. D. Procaccia and J. S. Rosenschein. Extensive-form argumentation games. In M. P. Gleizes, G. A. Kaminka, A. Now, S. Ossowski, K. Tuyls, and K. Verbeeck, editors, *EUMAS*, pages 312–322, 2005.

[19] I. Rahwan and K. Larson. Argumentation and game theory. In *Argumentation in AI*, pages 321–339. Springer, 2009.

[20] I. Rahwan, K. Larson, and F. Tohme. A characterisation of strategy-proofness for grounded argumentation semantics. In *Proceedings of the 21st IJCAI*, 2009.

[21] I. Rahwan, S. Ramchurn, N. Jennings, P. McBurney, S. Parsons, and E. Sonenberg. Argumentation-based negotiation. *Knowledge Engineering Review*, (18(4)):343–375, 2003.

[22] S. D. Ramchurn, N. R. Jennings, and C. Sierra. Persuasive negotiation for autonomous agents: A rhetorical approach. In *IJCAI Workshop on Comp. Models of Natural Argument, Acapulco, Mexico*, 2003.

[23] R. Riveret, H. Prakken, A. Rotolo, and G. Sartor. Heuristics in argumentation: A game theory investigation. In P. Besnard, S. Doutre, and A. Hunter, editors, *COMMA*, volume 172 of *Frontiers in Artificial Intelligence and Applications*, pages 324–335. IOS Press, 2008.

[24] D. Ross. Game theory. *The Stanford Encyclopedia of Philosophy (Winter 2012 Edition)*, 2012.

[25] D. Walton and E. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. SUNY series in Logic and Language. State University of New York Press, 1995.

[26] D. N. Walton. *Logical Dialogue-games And Fallacies*.

[27] S. Wells and C. A. Reed. A domain specific language for describing diverse systems of dialogue. *J. Applied Logic*, 10(4):309–329, 2012.

[28] T. Yuan, D. Moore, and A. Grierson. A conversational agent system as a test-bed to study the philosophical model DC. In *Proceedings of CMNA'03*, 2003.