Providing Arguments in Discussions Based on the Prediction of Human Argumentative Behavior*

Ariel Rosenfeld and Sarit Kraus
Department of Computer Science
Bar-Ilan University, Ramat-Gan, Israel 92500
rosenfa5@cs.biu.ac.il , sarit@cs.biu.ac.il

Abstract

Argumentative discussion is a highly demanding task. In order to help people in such situations, this paper provides an innovative methodology for developing an agent that can support people in argumentative discussions by proposing possible arguments to them. By analyzing more than 130 human discussions and 140 questionnaires, answered by people, we show that the well-established Argumentation Theory is not a good predictor of people's choice of arguments. Then, we present a model that has 76% accuracy when predicting peoples top three argument choices given a partial deliberation. We present the Predictive and Relevance based Heuristic agent (PRH), which uses this model with a heuristic that estimates the relevance of possible arguments to the last argument given in order to propose possible arguments. Through extensive human studies with over 200 human subjects, we show that peoples satisfaction from the PRH agent is significantly higher than from other agents that propose arguments based on Argumentation Theory, predict arguments without the heuristics or only the heuristics. People also use the PRH agent's proposed arguments significantly more often than those proposed by the other agents.

Introduction

Dialog, especially of an argumentative nature, is a highly demanding task for humans, both mentally and emotionally, as shown in discursive psychology research (Edwards 1997; Krauss 2001). Creative, analytical and practical abilities are needed to persuade or convince another person, (Sternberg 2008). An automated agent can help a human when engaging in an argumentative discussion by utilizing its knowledge and computational advantage to provide arguments to her.

When suggesting an argument to a human user, an agent can consider two possible approaches. First, the agent can suggest an argument that the person has (probably) considered and is prone to use anyway. Second, it can suggest an innovative argument that the person has (probably) not considered. In order to differ between the two approaches we need to asses which arguments people are likely to use given the current state of the discussion. First, we examine the well-established Argumentation Theory (see (Walton 2009) for an excellent summary) and its abilities to predict people's arguments. To date, very little investigation has been done regarding how well the proposed theories describe human reasoning. In this work, we present three experimental settings, with over 130 human conversations and 140 questionnaires, varying in complexity, which show the lack of predictive power of the existing Argumentation Theory. We introduce the concept of Bounded Rationality (Gigerenzer and Selten 2002) to the Argumentation Theory using the heuristics of *Relevance* and show that this simple amendment can provide the Argumentation Theory enriched predictive abilities. Second, we use Machine Learning (ML) techniques to provide a probability distribution over all known arguments given a partial deliberation. That is, our ML techniques provide the probability of each argument to be used next in a given discussion. Our model achieves 76% accuracy when predicting peoples top three argument choices given a partial deliberation. To construct our prediction model we utilize the psychological effect of confirmation bias (Nickerson 1998) and decision-making heuristics (Bonnefon et al. 2008).

Last, using the prediction model and the newly introduced heuristics of relevance, we designed and evaluated the Predictive and Relevance based Heuristic agent (PRH). Through extensive human studies with over 200 human subjects, we show that the PRH agent outperforms other agents that propose arguments based on Argumentation Theory, predicted arguments without the heuristics or only the heuristics on both axes we examined, i.e. peoples satisfaction from agents and people's use of the suggested arguments.

Related Work and Background

Argumentation Theory researchers have extensively studied the concept of a "good" argument and have proposed many theories explaining how to calculate these arguments (Walton 2009). Most of the proposed theories rely on some fundamental notions from (Dung 1995) and expand them in some way. These include the Bipolar Argumentation Framework (BAF) (Cayrol and Lagasquie-Schiex 2005b), the Value Argumentation Framework (VAF) (Bench-Capon

*We would like to thank Intel Collaboration Research Institute for Computational Intelligence for their support in this research.

2003) and the Weighted Argumentation Framework (WAF) (Dunne et al. 2011), to name a few, and a varity of applications such as ArgTrust (Tang, Sklar, and Parsons 2012a) and MIT's Deliberatorium (Klein 2011). This haystack of theories and applications is based on similar principles and ideas. It is common in Argumentation Theory to define some argumentation framework – a formalized structure in which statements (arguments) can attack or support each-other. Using different reasoning rules (semantics) it is possible to build sets of justified arguments (arguments that should be considered correct to some extent), and thereby solve the inherent conflicts.

It is argued that bipolarity in argumentation frameworks is essential to represent realistic knowledge (see (Amgoud et al. 2008) for a survey). Thus, throughout this work, we use the BAF modeling proposed in (Cayrol and Lagasquie-Schiex 2005b), denoted the "argumentation framework".

Definition. A Bipolar Argumentation Framework (BAF) $< A, R, S >$ consists of a finite set A called arguments and two binary relations on A called attack and support, respectively.

An argumentation framework can be represented as a directed graph with 2 types of edges[1]. A is the set of vertices, and R,S, are the sets of directed edges representing attack and support relations, respectively (see Figure 1 for illustration). In order to be able to perform reasoning about an argumentation framework, it is necessary to define reasoning rules, called *semantics*. Dung (Dung 1995) has defined several semantics which have been modified to fit the BAF model (Amgoud et al. 2008). In this work, we examine the 3 classical semantics proposed by Dung — Preferred, Grounded and Stable. Using the above semantics, a reasoner can identify sets of arguments, called *extensions*, which hold the special properties requested by the semantic. An argument which is a part of some extension is considered acceptable, valid or justified (to some extent).

Very little investigation has been done regarding how well the proposed models and semantics describe human reasoning. To the best of our knowledge only two papers address this topic; Rahwan et al. (Rahwan et al. 2010) studied the reinstatement principle in behavioral experiments and Cerutti et al. (Cerutti, Tintarev, and Oren 2014) examined humans' ability to evaluate formal arguments. These two works did not examine the possibility of the Argumentation Theory predicting people's argumentative behavior nor did they try to use their insights to generate advice for a user. Others studied the problem of which information should an agent reveal during a deliberation with people (Dsouza et al. 2013) or developed strategies for offers generation in human-agent negotiations (Rosenfeld et al. 2014). None of which did it in the context of Argumentation Theory.

The two approaches we examine when providing arguments to the user hold different rational-psychological explanations for why people would benefit from the suggested arguments. First, people search for *validation* for their existing opinions and beliefs (Linehan 1997). Thus, receiving consonant (supportive) suggestions to their views from an intelligent agent can help validate the person's beliefs. Second, Rational Choice Theory (Coleman and Fararo 1992) suggests that when an individual considers an action (e.g., argument to use) she needs to weigh all information that affect that action (argument). An agent can help a person by revealing additional information or help in weighing knowledge in an analytic manner.

It is common in literature to distinguish between different types of argumentation-structures (Walton et al. 2009). Throughout this work, we focus on *deliberations*, where the discussion process is aimed at exchanging opinions, beliefs and information and trying to reach some consensus on a controversial topic.

Predicting People's Argumentative Behavior

Data Collection

To obtain a better understanding of people's argumentative behavior we examined three experimental settings, varying in complexity, in which human subjects were asked to use arguments. Experiment 1 was a questionnaire based experiment where people were presented with a partial deliberation and were asked to choose, from a small list of 4 arguments, which argument they would use if they were one of the deliberating parties. Experiment 2 presents an analysis of real, free-form argumentative conversations. The subjects were not presented with a partial conversation and were not restricted in the manner they conduct the deliberation. Experiment 3 presents an analysis of semi-structured argumentative conversations, that is, the participants (deliberants) engaged in a deliberation while being restricted to arguments from a pre-defined list of 40 arguments. These settings allowed us to observe people's argumentative behavior in small and structured argumentative scenarios (Experiment 1), free-form argumentative deliberations (Experiment 2) and semi-structured argumentative conversations (Experiment 3).

Experiment 1 We requited 64 US citizens, all of whom work for Amazon Mechanical Turk (AMT), denoted the US-Group, ranging in age from 19 to 69 (mean=38, s.d.=13.7) with varying demographics, and a group of 78 Israeli Computer Science Bachelor students, denoted the IL-Group, ranging in age from 18 to 37 (mean=25, s.d.=3.7) with similar demographics, to take part in this experiment. Subjects were presented 6 fictional scenarios based on scenarios from (Walton 2005; Arvapally and Liu 2012; Cayrol and Lagasquie-Schiex 2005b; Amgoud et al. 2008; Tang, Sklar, and Parsons 2012b). Small changes were made in the original formulation of the scenarios to keep the frameworks small (6 arguments) and simply phrased, yet the scenarios were kept as close as possible to the origin. Each scenario was presented as a short conversation between 2 deliberants and the participant had to choose which of the 4 possible arguments he or she would use next if she was one of the deliberants. The following example is one of the 6 scenarios we presented to the subjects:

Example. *A couple is discussing whether or not to buy an SUV. Spouse number 1 (S_1): "We should buy an SUV; it's the right choice for us". Spouse number 2 (S_2): "But we*

[1] Sometimes called a bipolar interaction graph.

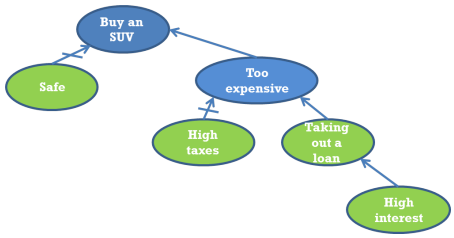

Figure 1: BAF: nodes are arguments. Arrows indicate attacks and arrows with diagonal lines indicate support.

can't afford an SUV, it's too expensive".
The subject was then asked to put himself in S_1's place and choose the next argument to use in the deliberation. The options were A. "Good car loan programs are available from a bank", B. "The interest rates on car loans will be high", C. "SUVs are very safe, safety is very important to us", D. "There are high taxes on SUVs".

The 6 scenarios were similar in the way they were presented; a short conversation of 2 statements and 4 possible argument from which to select. However, the argumentative structure they induced was different in order to simulate different argumentative complexity levels. Figure 1 presents a graphical representation of the example. This graphical representation was *not* presented to the subjects.

Experiment 2 We used real argumentative conversations form Penn Treebank Corpus (1995) (Marcus, Marcinkiewicz, and Santorini 1993) of transcribed telephone calls. The Penn Treebank Corpus consists of transcribed phone calls on various topics, among them some controversial topics such as "Should the death penalty be implemented?" and "Should a trial sentencing be decided by a judge or jury?", with which we chose to begin. We reviewed the 33 deliberations on "Capital Punishment" and 31 deliberations on "Trial by Jury" to identify the arguments used and cleared all irrelevant sentences (i.e, greetings, unrelated talk etc.). The shortest deliberations consisted of 4 arguments and the longest comprised 15 arguments (a mean of 7).

Experiment 3 In this experiment both subjects are aware of all arguments in the framework (perfect information). We restricted the interaction process by allowing deliberants to communicate only by using arguments from a pre-defined argument-list. We chose the topic of *"Would you get an influenza vaccination this winter?"* and constructed a pre-defined argument-list consisting of Pro (20) and Con (20) arguments. These arguments were extracted from debate sites[2] and medical columns[3]. The arguments were presented to the user alongside a "Discourse statements list", in which we provided a set of discourse markers and short statements such as "I agree", "I think that", "However" and others for subjects to use. We collected 72 deliberations between 144 Israeli college students, ranging in age from 21 to 39 (average of 29). The deliberations ranged in length from 5 arguments to 30 (mean 14).

[2]such as http://www.debate.org/ , http://idebate.org/

[3]such as http://healthresearchfunding.org/pros-cons-flu-shots/

Using Argumentation Theory to predict people's arguments

People's decision-making process is affected by a multitude of social and psychological factors and they often do not maximize expected utilities or use equilibrium strategies (Camerer 2003; Rosenfeld et al. 2012). The question we investigated in this work is whether, in the context of argumentative conversations, people would choose justified arguments according to some semantic choice. That is, would the Argumentation Theory provide predictive tools to predict human argumentative behavior.

Experiment 1 The results of experiment 1 were very surprising. For example, in the SUV scenario, most people (72%) chose the "closest" arguments in the framework - "Taking out a loan" and "High taxes" arguments, directly related to the last argument, where the "Taking out a loan" argument was the most popular one (37%). The "Taking out a loan" argument is supposed to be considered much weaker than the other 3 possible arguments by any classical semantics. The 3 leaf arguments should be considered strong and justified (as they are unattacked), whereas "Taking out a loan" is attacked by a strong argument. Such phenomena were encountered in all other scenarios as well; on average, a justified argument was selected only 67.3% of the time (under one of the classic semantics). Moreover, only 8% of the subjects chose justified arguments in all of the 6 scenarios[4].

Experiment 2 To analyze the conversations we constructed an argumentation framework for each topic, similar to the one in Figure 1, using the arguments we encountered in the extracted conversations from the Penn Treebank Corpus. The framework for "Capital punishment" consisted of 30 arguments and the framework for "Trial by Jury" consisted of 20 arguments. We calculated the classical Preferred, Grounded and Stable semantics (Cayrol and Lagasquie-Schiex 2005b) for the resulting 2 frameworks. Less than 45% of the arguments used by the subjects were part of some extension (under one of the classical semantics), with Preferred, Grounded and Stable semantics preforming very similarly (42%, 43%, 41%). It is important to note that 40% of the arguments in "Capital punishment" and 50% of the arguments in "Trial by Jury" were justified (under one of the classical semantics). That is, subjects chose justified arguments no better than the random selection would.

Experiment 3 The pre-defined argument list was translated into an argumentation framework (similar to Figure 1). Similarly to Experiment 2, we calculated the Preferred, Grounded and Stable semantics and each conversation was divided into 2 argument sets, A_1 and A_2, i. e., the arguments used by deliberant 1 and 2 respectively. Less than 50% of the arguments used by the subjects were part of some extension (under one of the classical semantics), with Preferred, Grounded and Stable semantics again preforming very similarly (44%, 49%, 41%). It is important to note that 50% of the arguments were justified (under one of the classical

[4]At least one justified argument existed in every scenario.

semantics). That is, the classical semantics were unable to indicate which arguments people would choose.

A common theme in the above 3 experiments is that people chose arguments with very little (if any) correlation with the arguments suggested by the classical semantics of the Argumentation Theory. Thus, we cannot rely solely on the theory in predicting people's arguments.

Using Machine Learning to predict people's arguments

We suggest a calculation of a measurements vector m, for each argument in an argumentation framework. This vector describes the argument and the context in which it is judged (the context in which a reasoner evaluates the argument). We divide m into 3 categories; *Justification* measurements, *Relevance Heuristic* values and *Confirmation Factor*.

Justification Dungs introduction of various extension-based semantics (Dung 1995) had a profound effect on the analysis of justified arguments. This idea was extended by Cayrol et al. (Cayrol and Lagasquie-Schiex 2005b; Amgoud et al. 2008) for BAFs. Unfortunately, as stated above, the mentioned semantics fail to predict the results of people's behavior in our experiments.

There have been a number of proposals for more sophisticated modeling and analysis of conflicting information, mainly incorporating relative strength / justification / credibility of the arguments. One commonly used proposal is the gradual valuation (Cayrol and Lagasquie-Schiex 2005a) in BAFs, denoted "Cayrol's calculation". The idea is to evaluate the *strength* of argument a using some aggregation function that conciliates between its attacking arguments' strength and its supporting arguments' strength. This recursive calculation, allows us to aggregate the number of supporters and attackers through the argumentation framework and reach a strength value in the [-1,1] interval for each argument. The strength value represents the deliberant's ability to support that argument, and defend it against potential attacks. The higher the strength level, the easier it is to support and defend the argument, and the harder it is to attack it. In our SUV example in Figure 1, Cayrol's suggested instantiation of gradual valuation J (proposed in (Cayrol and Lagasquie-Schiex 2005a)) provides J("Safe")=J("High Taxes")=J("High interest")=0 and J("Taking out a loan")=-0.33. This shows why "Taking out a loan" should be considered weaker than the other 3 proposed arguments in the example.

In an empirical article (Bonnefon et al. 2008), the authors examined a similar prediction problem of predicting people's choice between 2 options (for example, going to movie A or movie B) based on pro and con (supportive and attacking) information relevant to the options at hand. The main and most relevant insight from their work is that we should not ignore the number of supporting/attacking arguments when predicting people's choices. First, to identify the relation between every pair of arguments we used the four General Argumentation Heuristic Rules (Klein et al. 2003). For example, if argument a attacks b which in turn attacks c then a (indirectly) supports c. The influential arguments of argument a (Liao and Huang 2013), are the arguments whose status may affect the status of a, that is – they support/attack directly or indirectly. For each argument we considered the number of supporters (direct and indirect), the number of attackers (direct and indirect) and the supporters' portion among the influential arguments. For example, in our SUV example – "High taxes", "Tacking out a loan" and "High interest" are all influential arguments for "Too expensive" (see Figure 1).

It is important to state in this context that all of the above measurements rely solely on the argumentation framework, and as such require only a single calculation of their value for each framework regardless of the deliberation.

Relevance Heuristics At a given point in a deliberation, not all arguments are necessarily relevant. For instance, the argument "Safe" in our example (Figure 1) seems to be irrelevant to the context of the discussion, since the focus is on economic concerns. First, in order to identify the "relevant" arguments, we propose several distance measurements, both directed and undirected, that heavily rely on the current deliberation state. These distance measurements will help us investigate how the *proximity* between arguments, as portrayed by the edge-distance in the argumentation framework, truly affects the course of a deliberation.

We define 4 relevance measurements, each of which captures a different aspect of proximity. In the definitions, a denotes a possible argument, a_l is the last given argument, a_c is the "closest" argument to a which was previously given (by edge-distance metric) and α presents a designated argument which suggests an option/action which is the focus of the discussion (in our example it is whether or not to "Buy an SUV"). The relevance measurements of a possible argument a can be summed up in the following 4 points:

1. Minimum un/directed paths' length from a to a_l.
2. Minimum un/directed paths' length from a to a_c.
3. Minimum directed paths' length from a to α.
4. Minimum of all/some of the above features.

When omitting redundant calculations in the 4th criteria (i.e., the minimum of the minimal directed and undirected paths to a_l), 15 distinct measures remain.

As our results show, the directed paths' length from a to a_l and from a to α were found to be very influential. Despite this fact, to date they have not received any attention in the literature on Argumentation Theory. In our example, S_2's argument is considered a_l, and α is S_1's argument ("Buy an SUV"). When we consider a as "Safe", its distance to a_c or α (in this case, they are the same) is 1, while its directed distance to a_l is undefined and the undirected distance is 3. If a is "Taking out a loan" then its distance to a_l and a_c is 1 whereas its distance to α is 2. Therefore, every argument was assigned its relevance heuristic values — its directed paths' length from a to a_l and from a to α.

Confirmation Factor Confirmation bias is a phenomenon in psychology wherein people have been shown to actively seek out and assign more weight to evidence that confirms their beliefs, and ignore or underweight evidence that could disconfirm their beliefs (Nickerson 1998). In argumentative situations people may selectively consider arguments

in order to reinforce their expectations and disregard arguments that support alternative possibilities or attack their own. Practically, each argument has a *Confirmation Factor* which depends on the affects the argument has on the player's previously stated arguments. The Confirmation factor can be positive if it supports the previously used arguments by the deliberant. On the other hand it can be negative if it attacks previously used arguments. If the relation is ambiguous (both positive and negative) or unknown, then it has a neutral confirmation factor.

Hitherto, we described the features the different arguments hold (m). Separately, using the m values, we will calculate the features used in our prediction model. i.e., *deliberation context features* and *deliberant features* in every given stage of the discussion.

Deliberation context features In order to predict which arguments a deliberant would say in a given context of the conversation we also need to model the current state of the deliberation. To that aim, during the deliberation we account for the last 2 arguments used by the deliberant equipped with the agent and the last 2 arguments used by the other deliberant. We also indicate which of the deliberants used the last argument.

Deliberant features To capture the deliberant's preferences in argument types we calculated aggregated values of the arguments the deliberant used. Namely, we analyzed the arguments she used and calculated the average justification value (both the average J values and the support portion values), the average relevance values and the portion of times a confirmatory argument was used (out of the number of times one was available). In addition, we hold a *proneness feature*, in the [0,1] interval, which indicates the person's incline toward accepting a specific position on the discussed issue. For example, in a deliberation on "Capital punishment" a value of 1 means I support the Capital Punishment, 0 means I oppose it, and the higher the value the stronger my incline to agree with it. This feature stems from the Dissonance Theory (Festinger 1962) which suggests that once committed to an alternative (knowingly or unknowingly), people prefer supportive (consonant) information compared to opposing (dissonant) information to avoid or reduce post decision-making conflicts. In order to calculate the proneness feature we distinguished between 2 cases. In cases where the deliberant explicitly expressed his opinion (e.g., "I'm pro the death penalty") the proneness value is simply 1 or 0 (pending on the opinion expressed). In cases where the deliberant's opinion was not explicitly declared, we had to assess the deliberant's position using her previously stated arguments. We calculated this estimation using the portion of supportive arguments to the discussed issue that the deliberant used during the conversation. That is, using only supportive arguments is the same as explicitly stating your opinion.

Experiment 1: First, we calculated the set of features described above for each possible argument in every scenario. Then, given a learning period of k scenarios, where $k = 1, 2, \ldots, 5$, we took $6 - k$ scenarios out of the set. For each subject we calculated the deliberant features according to the arguments she selected in the scenarios in the set. For example, when $k = 5$, we took out 1 scenario at a time and calculated the average justification/relevance/etc. of the subject's selections in the 5 remaining scenarios. Finally, we labeled each set with the actual selections of the subject in the scenarios we removed.

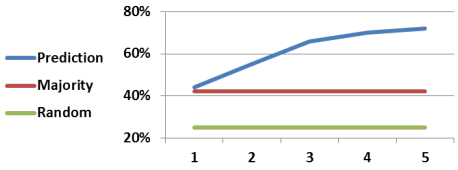

Figure 2: SVM's Learning Curve in Experiment 1.

We used 3 machine learning algorithms to test our feature set, i.e., the Support Vector Machine (SVM), Decision Tree Learning (DTL) and the Multi Layered Neural Network (MLNN). We trained and tested our model on the US-Group and the IL-Group separately (see the description in the Data Collection section). For both groups SVM was found to be the most accurate learning model of our observed data as it provided 72% and 78% accuracy in predicting the subject's 6th selection when learning from the first 5 (US-Group and IL-Group, respectively). DTL and MLMN both yielded less than 68% accuracy for both groups. For the US-Group, as the learning period (k) increased from 1 to 5, SVMs accuracy increased from 42% to 72%. That is, the more observations the prediction model had on the subjects' selections the higher its prediction accuracy. Random selection naturally provides 25% (in every scenario the subject was requested to choose 1 of 4 suggested arguments), and predicting the majority's selection (predicting the most popular selection among the other participants) provided 41% accuracy. See the learning curve in Figure 2. Interestingly, for both groups the features contributing to the prediction (using an entropy measurement) were (in the following order of importance):

1. Relevance (edge-distance from a to a_l).
2. Cayrol's justification calculation.
3. Support portion among the influential arguments.
4. Proneness.

Similar results were obtained for the IL-Group, wherein the prediction accuracy ranged from 45% (when $k = 1$) to 78% (when $k = 5$).

To check for cultural differences, we examined the use of the US-Group as a training-set and the IL-Group as the test-set. This setting achieved 76% accuracy, whereas using IL-Group as a training-set and the US-Group as a test-set demonstrated 69% accuracy. More surprising was the fact that the *very same features* were found to be influential in both settings. This may suggest that using a cross-cultural (US-Israeli) model is possible, though further investigation of this topic is needed.

Experiment 2: Similarly to Experiment 1, we first calculated the features describing the different arguments. Then, we randomly chose 5 deliberations on each deliberation topic (we used the 2 topics of "Capital Punishment" and "Trial by jury") and built 2 argumentation frameworks using the arguments given in those 5 deliberations. This resulted in

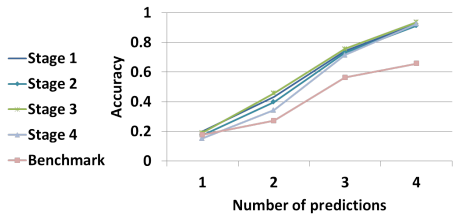

Figure 3: Prediction Curve for Capital Punishment.

2 frameworks comprising 30 arguments and 20 arguments, respectively.

Each conversation was then analyzed argument-by-argument (each argument is considered a stage/step in the deliberation). For each stage we created a vector containing the different deliberant features and the deliberation context features. The vector was then labeled with the argument which was used.

The remaining deliberations, which were not used to build the argumentation framework, were used for the training and test sets with the 1-left-out methodology. In other words, we learned from $n-1$ conversations and predicted the different stages of the 1-left-out conversation. The prediction was tested on varying starting stages, wherein we tested how the prediction quality changes over the time period of the deliberation in which we predict the arguments.

As a benchmark model we used the best model of 8 (simple) statistical models that essentially does not model the person but treats the argument selection process as a stochastic process. The *Bigram model* (Jelinek 1990) of the subject was found to be the best among the 8 models, using perplexity measurements. It outperformed the Trigram model of the subject as well as Bigram and Trigram statistical models of the other party in the deliberation. It also outperformed the combinations of the above.[5] Bigram modeling of the speaking deliberant calculates the probability $P(a_2|a_1)$ for every pair of arguments a_1, a_2. That is, the probability that a_2 follows a_1. These probabilities were estimated using a Maximum Likelihood Estimator on the data we collected. Given a_1 the model predicts $\text{argmax}_{a_2 \in A} P(a_2|a_1)$.

We trained and tested our model and found that DTL outperformed SVM and MLNN. Unlike in Experiment 1, in Experiment 2 there are many more arguments to consider in the prediction. Naturally, the prediction accuracy declined. However, if we use the probability measurements provided by the learning algorithm we can predict more than 1 argument – that is, we can predict the top x ranked arguments w.r.t their probability.On the topic of "Capital Punishment", in Figure 3, we can see how the prediction accuracy increases over the number of predicted arguments (X axis) and the stages from which we start our prediction (the different curves). When predicting the top 3 ranked arguments on the issue of "Capital Punishment" we achieved a prediction accuracy of 71%-76%, depending on the starting phase of the prediction. Very similar results were obtained for the "Trial by Jury" deliberation as well.

When comparing the influential attributes found in Experiment 1 and Experiment 2, we can see that the same features were found to be influential except for Cayrol's justification calculation that was ranked much lower in Experiment 2. The feature that indicated which deliberant used the last argument took its place. Note that this feature was not applicable in Experiment 1. The prediction accuracy sky-rocketed to 91.2% (Capital Punishment) and 88.6% (Trial by Jury) in cases in which the deliberant used more than one argument sequentially without interruption. In our study, 100% of the time, when a deliberant used more than one argument in a row, the second one was supportive and closely related to the first one. That is, the indication of which deliberant used the last argument was found to be very influential.

Regardless of the number of predictions, our models predictions reached better results than the benchmark model. To quantify this difference we used the MRR (Mean Reciprocal Rank) measure (Craswell 2009), which evaluates any process that produces a list of options ordered by their probability of correctness. Our model's MRR was 0.48 for Capital Punishment and 0.58 for Trial by Jury, whereas Bigram's MRR was 0.36 for both topics (the higher the better).

Agents for Arguments Provision

Policies

There are two main approaches when recommending an argument to a deliberant: recommend an argument that the deliberant has considered and would (probably) use anyway or recommend innovative arguments – those that the deliberant has (probably) not considered. We designed several policies which take these two approaches into consideration.

- **Predictive agent** (PRD), offers the top 3 ranked arguments in the prediction model. i.e., the arguments that fit the situation and the deliberant as learned from the training-set.
- **Relevance based heuristic agent** (REL), offers the 3 "closest" arguments to the last given argument (using edge-distance). As we previously observed, the relevance features were found to be very influential in the prediction model. Therefore, we wanted to test whether the relevance notion could act as a good policy, without any prediction or complex modeling.
- **Weak Relevance based heuristic agent** (WRL), offers the 3 least related arguments to the last argument (using edge-distance). The idea behind this policy is to offer the subject arguments that she would not naturally contemplate or say.
- **Predictive and Relevance based Heuristic agent** (PRH), offers the top 2 predicted arguments and the most relevant argument (using edge-metric) which was not part of the predicted arguments. This policy attempts to enjoy the best of the two policies.
- **Theory based agent** (TRY), which calculates the extension of the argumentation framework using Grounded semantics and offers 3 arguments which are part of that extension. Because the extension is usually larger than 3, we offer the 3 "closest" arguments to the last given one (using edge-distance). That is, among "justified" arguments, the agent offers the top 3 arguments relevant at the moment.

[5] All models used Backoff to avoid the assignment of 0s

- **Random agent** (RND), offers 3 arguments in a random fashion while avoiding previously used arguments. This policy served as a benchmark.

Experimental Evaluation

We used Experiment 3 conversations on "Influenza Vaccinations" to train our prediction model.

Second, we implemented the 6 different agents; each of them was tested in 17 chats, totaling 102 deliberations with 204 human subjects. In each chat we coupled 2 subjects who were asked to deliberate over the same topic of influenza vaccination, but in a *free form chat*. Only one subject in each couple was assigned a personal agent to maintain scientific integrity. All 204 subjects who took part in this stage were Israeli students who were recruited from classrooms, libraries, etc. The subject ranged in age from 18 to 30.

The identification of the arguments used by the deliberants was done in a *Wizard of Oz* fashion, where during the chat an expert from our lab mapped the given sentences into the known arguments in the previously built BAF (consisting of 40 arguments). The deliberant who was assigned an agent received 3 suggestions on the right side of the screen in a textual form, after each argument used (by either of the deliberants). Suggestions started to appear after encountering 2 arguments in the deliberation to enable a short learning period for the agent. We emphasize that the agent had no prior knowledge of the deliberant and required no information from the subject during the deliberation. Participants could not select a suggested argument by clicking on it, but had to type their arguments in a designated message-box. This restriction was implemented to avoid "lazy" selections.

All obtained deliberations consisted of 4-20 arguments (mean 9), and took between 5-21 minutes (mean 12). Deliberations ended when one of the deliberants chose to end it, just as in real life. Yet, in order to receive the 15 NIS payment (the price of a cup of coffee and a pastry in the University cafeteria), the deliberants had to deliberate for a minimum of 5 minutes.

At the end of each session, the subject who was equipped with an agent was asked to provide her subjective benefit from the agent on the following scale; Very positive, Positive, Neutral (neither positive nor negative), Negative, Very Negative.

Results We analyzed the 102 deliberations using the *Reported Benefit* and the *Normalized Acceptance Rate* which is defined as follows: For each conversation we calculated the percentage of arguments the subject used from the agent's suggestion. Then we averaged those percentages to calculate the *Normalized Acceptance Rate* for each agent. The Normalized Acceptance Rate of the PRH agent was significantly higher than the other agents, averaging 62% acceptance (the subject's acceptance rate ranged between 20% and 100%), whereas PRD averaged 26% (0%-50%) and REL averaged 47% (10%-100%). WRL, RND and TRY performed very poorly achieving 3%, 10% and 11%, respectively. This result was found to be significant in the $p < 0.05$ range using post-hoc univariate ANOVA. See Graph 4.

As for the *Reported Benefit*, again, the PRH agent outperformed the others in a convincing manner. All 17 subjects

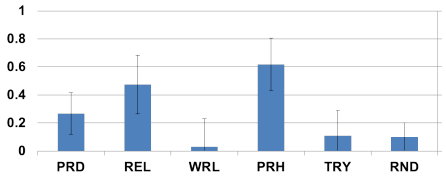

Figure 4: Normalized Acceptance Rate for the agents.

equipped with the PRH agent reported a positive benefit (5 reported very positive, 12 reported positive), which is significantly better than the contending agents with $p = 0.04$ using Fisher's Exact test. For comparison, PRD achieved 2 very positive benefits, 10 positive and 5 neutral benefits, whereas for REL no one reported very positive benefits, 12 reported positive and 5 reported neutral benefits. RND, WRL and TRY again performed very poorly with very few subjects reporting positive benefits.

Interestingly enough, no one reported negative or very negative benefits for any of the agents. The fact that no one reported a negative or a very negative benefit is very encouraging; namely, even when the advice was not used by the subject, the agent did not "bother" them. This finding indicates that argument provision agents, regardless of the algorithm, hold much potential in real world implementation.

Conclusions and Future Work

We performed an extensive empirical study, with over 400 human subjects and 250 annotated deliberations, on the prediction of argument selection and its use in designing suggestion policies for automated agents. We conclude that the incorporation of Machine Learning in argumentation is needed to investigate argumentation in the real world. Combining the *Relevance* notion, which was first introduced in this paper, with the abstract Argumentation Theory should provide additional predictive strength. Moreover, other aspects of argumentation in addition to justification should be explored to better gap the differences between human argumentative behavior and the Argumentation Theory. Even though the prediction model yields limited accuracy, using it provided solid policies.

Regardless of policy, no one reported a negative or a very negative benefit from the agent's suggestions. This finding emphasizes the potential automated agents hold in the context of argument suggestion during argumentative discussions.

During the research process we constructed an annotated corpus that we would be pleased to share for future research. We intend to expand this methodology and use the features and insights provided in this study to design and implement repeated-interaction agents. These agents could learn from past chats by the user, on different topics, and tailor a suggestion policy for her. As part of this work we will examine the exploration of different policies over time, the user modeling for multiple interactions and the ability to deduce insights from one conversation topic to another. Proceeding on different path, we would also like to explore how our agent could be adapted to help people in different argumentative forms such as negotiations.

References

Amgoud, L.; Cayrol, C.; Lagasquie-Schiex, M.-C.; and Livet, P. 2008. On bipolarity in argumentation frameworks. *International Journal of Intelligent Systems* 23(10):1062–1093.

Arvapally, R. S., and Liu, X. F. 2012. Analyzing credibility of arguments in a web-based intelligent argumentation system for collective decision support based on k-means clustering algorithm. *Knowledge Management Research & Practice* 10(4):326–341.

Bench-Capon, T. J. 2003. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation* 13(3):429–448.

Bonnefon, J.-F.; Dubois, D.; Fargier, H.; and Leblois, S. 2008. Qualitative heuristics for balancing the pros and cons. *Theory and Decision* 65(1):71–95.

Camerer, C. 2003. *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Cayrol, C., and Lagasquie-Schiex, M.-C. 2005a. Graduality in argumentation. *J. Artif. Intell. Res.(JAIR)* 23:245–297.

Cayrol, C., and Lagasquie-Schiex, M.-C. 2005b. On the acceptability of arguments in bipolar argumentation frameworks. In *Symbolic and quantitative approaches to reasoning with uncertainty*. Springer. 378–389.

Cerutti, F.; Tintarev, N.; and Oren, N. 2014. Formal arguments, preferences, and natural language interfaces to humans: an empirical evaluation. In *21st European Conference on Artificial Intelligence*.

Coleman, J. S., and Fararo, T. J. 1992. Rational choice theory. *Nueva York: Sage*.

Craswell, N. 2009. Mean reciprocal rank. In *Encyclopedia of Database Systems*. Springer. 1703.

Dsouza, S.; Gal, Y.; Pasquier, P.; Abdallah, S.; and Rahwan, I. 2013. Reasoning about goal revelation in human negotiation. *Intelligent Systems, IEEE* 28(2):74–80.

Dung, P. M. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence* 77(2):321–357.

Dunne, P. E.; Hunter, A.; McBurney, P.; Parsons, S.; and Wooldridge, M. 2011. Weighted argument systems: Basic definitions, algorithms, and complexity results. *Artificial Intelligence* 175(2):457–486.

Edwards, D. 1997. *Discourse and cognition*. Sage.

Festinger, L. 1962. *A theory of cognitive dissonance*, volume 2. Stanford university press.

Gigerenzer, G., and Selten, R. 2002. *Bounded rationality: The adaptive toolbox*. MIT Press.

Jelinek, F. 1990. Self-organized language modeling for speech recognition. *Readings in speech recognition* 450–506.

Klein, M.; Sayama, H.; Faratin, P.; and Bar-Yam, Y. 2003. The dynamics of collaborative design: insights from complex systems and negotiation research. *Concurrent Engineering* 11(3):201–209.

Klein, M. 2011. How to harvest collective wisdom on complex problems: An introduction to the mit deliberatorium. *Center for Collective Intelligence working paper*.

Krauss, R. M. 2001. The psychology of verbal communication. *International Encyclopaedia of the Social and Behavioural Sciences* 16161–16165.

Liao, B., and Huang, H. 2013. Partial semantics of argumentation: basic properties and empirical. *Journal of Logic and Computation* 23(3):541–562.

Linehan, M. M. 1997. *Validation and psychotherapy.* American Psychological Association.

Marcus, M. P.; Marcinkiewicz, M. A.; and Santorini, B. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics* 19(2):313–330.

Nickerson, R. S. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2(2):175.

Rahwan, I.; Madakkatel, M. I.; Bonnefon, J.-F.; Awan, R. N.; and Abdallah, S. 2010. Behavioral experiments for assessing the abstract argumentation semantics of reinstatement. *Cognitive Science* 34(8):1483–1502.

Rosenfeld, A.; Zuckerman, I.; Azaria, A.; and Kraus, S. 2012. Combining psychological models with machine learning to better predict peoples decisions. *Synthese* 189(1):81–93.

Rosenfeld, A.; Zuckerman, I.; Segal-Halevi, E.; Drein, O.; and Kraus, S. 2014. Negochat: A chat-based negotiation agent. In *AAMAS*.

Sternberg, R. 2008. *Cognitive psychology*. Cengage Learning.

Tang, Y.; Sklar, E.; and Parsons, S. 2012a. An argumentation engine: Argtrust. In *Ninth International Workshop on Argumentation in Multiagent Systems*.

Tang, Y.; Sklar, E.; and Parsons, S. 2012b. An argumentation engine: Argtrust. In *Ninth International Workshop on Argumentation in Multiagent Systems*.

Walton, D.; Atkinson, K.; Bench-Capon, T.; Wyner, A.; and Cartwright, D. 2009. Argumentation in the framework of deliberation dialogue. *Argumentation and Global Governance*.

Walton, D. N. 2005. *Argumentation methods for artificial intelligence in law*. Springer.

Walton, D. 2009. Argumentation theory: A very short introduction. In *Argumentation in artificial intelligence*. Springer. 1–22.