The Pragmatic Parliament: A Framework for Socially-Appropriate Utterance Selection in Artificial Agents

Felix Gervits (felix.gervits@tufts.edu)
Human-Robot Interaction Laboratory, Tufts University, Medford, MA 02155

Gordon Briggs (gordon.briggs.ctr@nrl.navy.mil)
NRC Postdoctoral Fellow, Naval Research Laboratory, Washington, DC 20375

Matthias Scheutz (matthias.scheutz@tufts.edu)
Human-Robot Interaction Laboratory, Tufts University, Medford, MA 02155

Abstract

One of the hallmarks of human natural language (NL) interaction is the ability for people to balance a variety of social and communicative goals when choosing how to realize their speech actions. These goals can include pragmatic criteria such as correctness, informativeness, and brevity (i.e., Gricean conversational maxims) or social factors such as politeness. However, there currently does not exist a general algorithmic method to explicitly modulate language generated by artificial agents based on an arbitrary number of pragmatic and social criteria. We propose a novel method to accomplish this task, in which rankings of candidate utterances by different pragmatic or social criteria are fused by use of a voting algorithm. We then give a proof-of-concept demonstration of the application of this method in the context of directive generation for human-robot interaction.

Keywords: Human-Robot Interaction; Pragmatics; Natural Language Generation; Politeness

Introduction

One of the key strengths of humans as social agents is the ability to adapt our language to the communicative norms and needs of the present situation. When giving directives and making requests, we know when it is appropriate to be terse and direct (e.g., “Move out, double-time!”), and when it is appropriate to be polite and circumspect (e.g., “Would you mind passing the salt, please?”). In all our natural language (NL) interactions, we are faced not only with the complex problem of what to say, but also how to say it. Much of this complexity originates from the fact that the intended meaning of utterances in different situational contexts often differs with the literal meaning. For example, asking a waiter, “Can I have a steak?” is not a literal query as to one’s physical ability to possess a particular menu item, but rather a means to convey an order.

Dialogue interaction for artificial agents is often viewed from a plan-oriented standpoint, in which the key plan-operators are speech actions used to achieve some high-level set of task goals. The precise way in which these speech actions are realized (in so far as it does not affect the efficacy of the speech act) is often of secondary concern. As NL-enabled agents become more prevalent in society, and as their manufacturers increasingly market these devices as “social” agents[1], the disparity between the state-of-the-art in computational NL systems and the richness of human-generated language will become increasingly apparent. As such, the ability for an NL-enabled agent to consider and modulate their generated language in human-like ways will become correspondingly more relevant and important.

There is a sizable literature that draws inspiration from pragmatics and socio-linguistics in order to address specific subproblems in natural language generation (NLG) at the subsentential, sentential, and discourse levels. For example, there has been extensive work in operationalizing Gricean pragmatic criteria (Grice, 1975) at the subsentential level, specifically in the area of referring expression (RE) generation (Dale & Reiter, 1995; Krahmer & Van Deemter, 2012), in which considerations of correctness, informativeness, and brevity are addressed. There also exists a small body of work that seeks to modulate NLG at the sentential level (Briggs & Scheutz, 2013; Gupta, Walker, & Romano, 2007; Miller, Wu, & Funk, 2008). These approaches seek to operationalize the notion of *face-threat* from politeness theory, and adjust the behavior of an agent accordingly.

Much of the previous work at the intersection of pragmatics, socio-linguistics, and NLG focuses on tackling specific subproblems in NLG or on modulating language based on a small set of criteria, such as politeness, e.g., Gupta et al. (2007). Yet, in order to generate more human-like language, a much more general framework is necessary. Below we propose some features that such a framework should possess:

1. The method of NLG modulation should be able to explicitly consider an *extensible* number of pragmatic and socio-linguistic criteria.

2. The method of NLG modulation should be *adaptable* such that the current situational context may affect the relative importance of communicative criteria.

3. The method of weighing communicative criteria should be *agnostic* to the choice of the underlying semantic representations used by the system.

At present, there exists no framework that meets all of these criteria. Much of the work in RE generation implicitly considers pragmatic criteria in the design of its algorithms (i.e., RE generation algorithms often search in order of shortest to

[1] e.g., JIBO: http://www.jibo.com

longest solution and terminate when a sufficiently informative solution is found (Bohnet & Dale, 2005)), but does not provide an extensible framework for pragmatic and sociolinguistic modulation. Work such as Briggs and Scheutz (2013) is extensible, but it sorts potential utterances according to a fixed preference ordering of communicative goals, and its adaptability is limited. The work in Bayesian cognitive models of pragmatics (Goodman & Stuhlmüller, 2013) can be extended to account for social communicative criteria, but it is tightly coupled to semantic representations and small domains amenable to Bayesian computational algorithms. Finally, there are promising approaches which meet some of the requirements, but they are limited to specific domains such as tutoring (e.g., Moore, Porayska-Pomsta, Varges, and Zinn (2004); Nye, Graesser, and Hu (2014)), and do not offer general solutions outside of that context.

In the following section, we present an approach that possesses all of the above desired features. We focus, in this paper, on the problem of modulating generated language at the *sentential* level, though we hope to apply similar techniques to NLG problems at subsentential and discourse levels. We first begin by examining various communicative goals that NL-enabled agents may need to consider. Next, we present a novel method of balancing these communicative criteria based on techniques from social choice theory (specifically, voting algorithms). Finally, we demonstrate our approach in the context of a human-robot interaction (HRI) scenario, and discuss directions for future work.

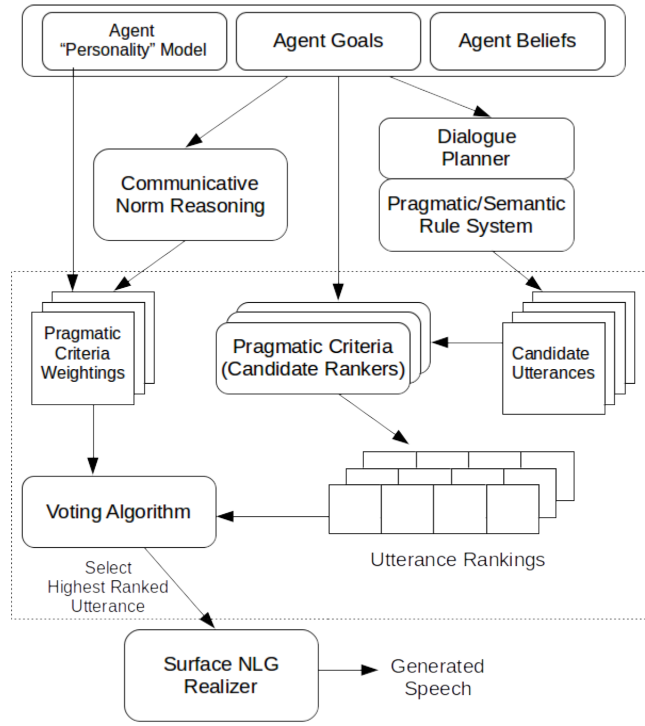

Figure 1: Diagram outlining an architecture for flexible NLG that is modulated by an extensible number of pragmatic criteria. The dotted line represents the architectural components we focus in detail on in this paper.

Utterance Selection

In this section we describe an utterance selection algorithm designed to achieve the sort of linguistic modulation we have proposed. In Figure 1 we outline the key components to this approach, which bridges, within the context of an NLG architecture, the output of a dialogue planning component (responsible for selecting an appropriate sequence of speech actions to achieve some agent goal) and the input of an NLG surface realizer component, which is responsible for translating some symbolic linguistic representation into text to be displayed or to be output via text-to-speech. In many architectures, this connection is direct. However, as we have previously addressed, there are multiple ways of realizing speech actions. To effectively consider them, we need the following components:

- A component that factors situational context to produce multiple potential candidate utterance realizations for a given speech action. Examples of NLG pipelines that include such a component are Briggs and Scheutz (2013) and Gupta et al. (2007).

- A set of pragmatic or social criteria $\mathbf{P}$, each with a corresponding utility function U_ρ ($\rho \in \mathbf{P}$), that generates a weak preference order over candidate utterances (Υ). These criteria include *correctness* (Maxim of Quality), *informativeness* (Maxim of Quantity), *directness* and *brevity* (Maxim of Manner), and *politeness*.

- A component that factors in the agent's beliefs about the current situational context, current goals, and potentially any "personality" model given to the agent in order to produce a set of weights for each pragmatic criterion: $\mathbf{W} = \{W_1, \ldots, W_{|\mathbf{P}|}\}$, where $W_\rho \in \mathbb{N}$ denotes the current strength of criteria ρ.

- A component that merges the rankings of candidate utterances Υ produced by the pragmatic criteria evaluations $(U_1, \ldots, U_{|\mathbf{P}|})$ in accordance with the weights generated by the communicative norm reasoner.

In order to merge the rankings of candidate utterances, we used the Schulze voting method (Schulze, 2011), where each ordering produced by U_ρ was counted W_ρ times. This voting method is a ranked single-winner election system from social choice theory, which is used by many organizations to select a candidate that maintains voters' individual preferences. While this approach has not been previously applied to the domain of computational pragmatics, we find that it offers a robust, computationally-tractable solution to the problem of balancing communicative goals in natural language generation. In the following sections, we present a proof of concept demonstration of our framework, and show how it can be used to generate socially-appropriate directives in the context of human-robot interaction.

Table 1: Utterance selections for various communicative criteria priority orderings

Relative Criteria Weightings	Utterance Selected	Utterance Output
Directness >Brevity >Politeness	*Instruct(R,β,do(β,plug_in(R)),{})*	"Plug me in"
Directness >Politeness = Brevity	*Instruct(R,β,do(β,plug_in(R)),{please})*	"Plug me in"/"Plug me in, please"
Brevity >Politeness >Directness	*AskYN(R,β,capableOf(β,plug_in(R)),{})*	"Could you plug me in?"
Politeness >Brevity >Directness	*AskYN(R,β,capableOf(β,plug_in(R)),{please})*	"Could you plug me in, please?"
Directness = Politeness = Brevity	*Instruct(R,β,do(β,plug_in(R)),{})*	"Plug me in"/"Plug me in, please"
Politeness >Directness = Brevity	*AskYN(R,β,capableOf(β,plug_in(R)),{please})*	"Could you plug me in, please?"

Demonstration: Directive Generation

In order to demonstrate the generality of this framework, we describe how our proposed framework has been integrated with the NL pipeline in a cognitive, robotic architecture, DIARC (Schermerhorn, Kramer, Middendorff, & Scheutz, 2006). There has been growing interest in the field of HRI in the ways in which robots could phrase requests for assistance from human interaction partners with respect to politeness and other social norms (Gupta et al., 2007; Srinivasan & Takayama, 2016; Strait, Canning, & Scheutz, 2014; Torrey, Fussell, & Kiesler, 2013). Below we present how our framework can be used to address this challenge.

Framework Configuration

In DIARC, utterances are represented in the following form:

$$\upsilon = UtteranceType(\alpha, \beta, X, M)$$

where *UtteranceType* denotes the speech act classification, α denotes the speaker, β denotes the addressee, X denotes an initial semantic analysis, while M denotes a set of sentential modifiers (e.g., "please"). The pragmatic reasoning component in the architecture associates an utterance υ in context C with a set of implications:

$$\upsilon_C := \langle \mathbf{B_{lit}}, \mathbf{B_{int}}, \theta \rangle$$

Each rule associates a particular utterance form υ in context C with a tuple containing the set of beliefs $\mathbf{B_{int}}$ to be inferred based on the intended meaning of the utterance, the set of beliefs to be inferred based on the literal meaning of the utterance $\mathbf{B_{lit}}$, as well as the degree θ to which the utterance can be considered a face-threatening act (i.e., a threat to a person's self-image or autonomy) in context C (Brown & Levinson, 1987).

We define the criterion of *correctness* as:

$$U_{correct}(\upsilon_C, \beta) = -|\{x : x \in \mathbf{B_{int}}(\upsilon_C) \wedge \beta \not\vdash x\}|$$

where β consists of the agent's current set of beliefs. Therefore, utterances that imply more facts unsupported by the agent's beliefs are considered less correct than those that imply fewer unsupported facts. We define the criterion of *informativeness* as:

$$U_{inform}(\upsilon_C) = |\mathbf{B_{int}}(\upsilon_C)|$$

such that utterances that imply more facts are considered more informative than those that imply fewer facts. We define the criterion of *directness* as:

$$U_{direct}(\upsilon_C) = \begin{cases} 1 & \mathbf{B_{lit}} = \mathbf{B_{int}} \\ 0 & \mathbf{B_{lit}} \neq \mathbf{B_{int}} \end{cases}$$

such that utterances in which the literal and intended meanings are the same are considered more direct than those in which they differ. We define the criterion of *politeness* as:

$$U_{polite}(\upsilon_C) = -\theta(\upsilon_C)$$

such that utterances in which the associated face-threat value (θ) are lower are considered more polite than those in which in it is higher. Finally, we define the criterion of *modifier-brevity* such that:

$$U_{m-brevity}(\upsilon_C) = -|M|$$

utterances with fewer sentential modifiers are considered briefer than those with more sentential modifiers[2].

Example Scenario

In this section, we present a proof-of-concept demonstration of the pragmatic modulation framework as applied to a directive formulation problem. Consider a scenario in which an NL-enabled robot is low on charge and needs a human to plug it in ($want(bob, plug\_in(self))$). This will require a directive to be formulated and communicated to the human in order to accomplish the end goal of being plugged in. We consider four main directive formulation strategies in this scenario, realized in the following pragmatic rules in the architecture's dialogue component[3]:

$$Instruct(\alpha, \beta, X, M) := \langle \{want(\alpha, bel(\beta, want(\alpha, X)))\}, \{want(\alpha, bel(\beta, want(\alpha, X)))\}, \theta_{instruct} \rangle \quad (1)$$

represents a literal directive from α to β. In the case of no politeness softeners, $M = \emptyset$, where in the case of softeners,

[2]Ideally, an operationalization of brevity should obtain some metric from the surface realization of a potential utterance (e.g., phoneme count, simulated speech output time, etc.). This architectural integration is still a work in progress.

[3]While DIARC has the capacity to handle unconventional indirect requests (e.g., "My batteries are running low..."), for sake of clarity we focused on more conventional cases in our demonstration.

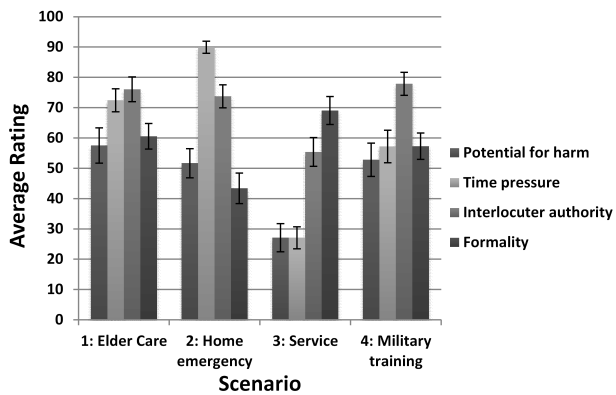

Figure 2: Ratings of social context dimensions from behavioral data. Error bars represent SEM.

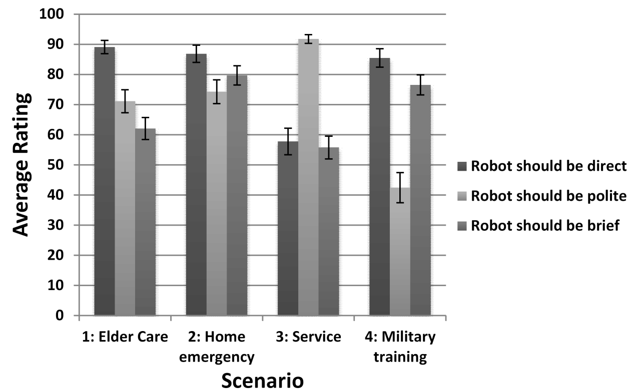

Figure 3: Ratings of pragmatic criteria from behavioral data. Error bars represent SEM.

$M = \{please\}$. In contrast, an indirect request can be represented by:

$$\begin{aligned} &AskYN(\alpha, \beta, capableOf(\beta, X), M) := \\ &\quad \langle \{want(\alpha, informif(\beta, \alpha, capableOf(\beta, X)))\}, \\ &\qquad \{want(\alpha, bel(\beta, want(\alpha, X)))\}, \theta_{AskYN} \rangle \end{aligned} \tag{2}$$

which represents the query "Can you X?" It is literally a query about one's capability, but can be interpreted as an indirect request. In the case of no politeness softeners, $M = \emptyset$, where in the case of softeners, $M = \{please\}$. The relative face-threat values for each strategy are: $\theta_{AskYN-p} < \theta_{AskYN} < \theta_{instruct-p} < \theta_{instruct}$, where "p" indicates the presence of politeness softeners.

Table 1 contains the utterance forms selected by the voting algorithm given the relative weights of the communicative goals of directness, politeness, and brevity. Correctness and informativeness were weighted above these criteria, but for the purposes of this scenario were irrelevant (as all candidate utterances were equally correct and informative). Our framework allows for socially-appropriate directive generation, as the various directive strategies, including: *Direct* - "Plug me in", *Direct with softener* - "Plug me in, please", *Indirect* - "Could you plug me in?", and *Indirect with softener* - "Could you plug me in please?" were generated in different potential contexts. For example, if directness is the top priority (e.g., in a task-oriented environment) then a direct utterance will be selected. However, if politeness is required (e.g., in casual conversation or a service-robot scenario) then a more indirect utterance will be selected. The results of the demonstration show how our framework can be integrated in a dialogue system in order to produce robust socially-sensitive natural language utterances in a variety of contexts.

Setting the Pragmatic Criteria Weightings

Next, we conducted an empirical investigation to establish an initial set of weights for the model (see 'Pragmatic Criteria Weightings' component in Figure 1) that is consistent with human judgments. To this end. we conducted a crowd sourcing study on Amazon Mechanical Turk in which people were shown hypothetical human-robot interactions and asked to rate various features of the interactions. A total of 42 people participated in the study - 23 of the participants were male, 17 were female, and 2 did not specify a gender. The average age was 35.9. All participants had US zip codes and received $1 for their participation. The study was approved by the Tufts Institutional Review Board and all participants gave informed consent. In the study, participants were shown a text description of four scenarios[4] and were asked to rate various social context dimensions (potential for harm, time pressure, interlocuter authority, and formality) and pragmatic criteria (directness, politeness, brevity) associated with the robot's speech in each scenario on a sliding scale from 0 (Strongly Disagree) to 100 (Strongly Agree).

Analyses of the data were carried out in order to establish a mapping between the pragmatic criteria, weightings, and utterance selection. First, the results for social context dimensions (see Figure 2) showed that each scenario had a distinct feature profile. Consequently, people expected the robot to adopt a different set of pragmatic criteria in each scenario (see Figure 3). The link between these contextual dimensions and the corresponding pragmatic criteria is important for determining the model weights in new contexts, but this will require future investigations that address the problem directly (see Discussion section). For the present work, we focus on using people's ratings for the pragmatic criteria to set the initial weights of our model. In order to rank these weights, we conducted a repeated measures ANOVA (with Bonferroni

[4]Scenario 1 involved an elder care setting in which a robot asked the nurse for a sick patient's medication ("Hand me the red pills."). Scenario 2 involved a household robot running low on battery that asked to be plugged in before important data was lost ("Plug me in."). Scenario 3 involved a service robot that requested to take a child's coat at a fancy reception ("Hand me your coat."). Finally, Scenario 4 involved a mine-sweeping robot that asked its superior officer to step aside as it searched a room in a training exercise ("Move out of the way.").

Table 2: Candidate utterance types with corresponding directives from Scenario #2

Utterance Type	Robot Directive
(u_1) Direct	"Plug me in."
(u_2) Direct with softener	"Plug me in, please"
(u_3) Indirect statement	"I would like you to plug me in.
(u_4) Indirect statement with softener	"I would like you to plug me in, please."
(u_5) Indirect question	"Could you plug me in?"
(u_6) Indirect question with softener	"Could you plug me in, please?"

correction) to tease out the ordering of the pragmatic criteria for each scenario. In scenario 1 ($F(2,82) = 18.237$, $p < .001$), post-hoc tests revealed that people expected the robot to be more direct (89%) vs polite (71%, $p < .005$) and brief (62%, $p < .005$). There was no significant difference between politeness and brevity in this scenario ($p = .309$). This corresponds to criteria weightings of Direct > Polite = Brief, which would result in a tie in the selected utterance: "Hand me the red pills"/"Hand me the red pills, please" (see Table 1). In scenario 2 ($F(2,82) = 4.470$, $p < .05$), post-hoc tests revealed that people expected the robot to be slightly more direct (87%) vs polite (74%, $p < .05$). However, there was no significant difference between directness and brevity in this scenario ($p = .092$) or between politeness and brevity ($p = .673$). This corresponds to criteria weightings of Direct = Polite = Brief, and a tie in the selected utterance: "Plug me in"/"Plug me in, please". In scenario 3 ($F(2,82) = 44.334$, $p < .001$), post-hoc tests revealed that people expected the robot to be more polite (92%) vs direct (58%, $p < .001$) and brief (56%, $p < .001$). There was no significant difference between directness and brevity in this scenario ($p = 1.00$). This corresponds to criteria weightings of Polite > Direct = Brief, and a selected utterance of "Could you hand me your coat, please". Finally, in scenario 4 ($F(2,82) = 32.004$, $p < .001$), post-hoc tests revealed that people expected the robot to be more direct (85%) vs polite (42%, $p < .001$) and brief (77%, $p < .005$). People also expected the robot to be more brief vs polite ($p < .001$). This corresponds to criteria weightings of Direct > Brief > Polite, and a selected utterance of "Move out of the way". The utterance output corresponding to each of these criteria weightings is listed in Table 1, and was selected from a list of 6 possible utterance types (see Table 2). Overall, these empirical results serve as a starting point by which to set the weights of our model for socially-appropriate utterance selection. Extensions of this approach as well as suggestions for future work are discussed in the Discussion below.

Discussion

In the previous section, we demonstrated how the application of our novel, pragmatically-sensitive framework can result in richer, more human-like modulation of NL. The method of explicit operationalization of pragmatic and socio-linguistic criteria into functions that can produce preference orderings over candidate NLG representations holds advantages over many of the pre-existing approaches. For example, the merging of preference orders produced by utility functions rather than the direct merging of utility values avoids tricky questions about the direct quantitative comparisons of different pragmatic and socio-linguistic criteria [5]. Additionally, the explicit operationalization of criteria allows for more extensibility and flexibility compared to algorithms in which communicative criteria are factored in implicitly. Nonetheless, this extensibility and flexibility leads to a variety of challenges for future work.

Computing and Learning Criteria Weights

While we used an empirical approach to initially set the weights for utterance selection, there still exists the normative challenge of determining what the most appropriate orderings/weightings of pragmatic and social goals are in any given communicative context. We allude to possible sources of information that could be used to compute these weights in Figure 1. These include the current beliefs of the agent about the situational context, the agent's goals (task-goals and social-goals), and potentially even models of personality (Mairesse & Walker, 2011) or culture (Endrass & André, 2014) that a designer may wish to imbue in the agent (e.g., a social robot configured to be impolite for entertainment purposes). The *dynamics* of how weights change within a single interaction and context are also a matter for future investigation. For example, a robot could become more polite if it detects that its interlocutor is distressed. The appropriate solution for this component would be entirely dependent on the particular interaction purpose, context, and desired effect. We view the present work as the first necessary step to opening up this rich area of future research.

We envision the process of computing criteria weights as a two-step process. First, various observable or inferable social context factors are evaluated in the given interaction scenario. These contextual features may include factors such as those in Figure 2. These in turn govern the weights that modulate utterance selection. The mapping between social context features and communicative criteria weights could potentially be learned in the following ways. *Explicit feedback*: the human interactant could provide explicit negative or positive feedback about the agent's recently-produced utterance with respect to a particular communicative criterion (e.g., "That was rude!" would indicate that weights for politeness should be increased in the present context). More subtle cues from facial expression, body language, or affect could also be used to modulate politeness, as in Moore et al. (2004). *Passive observation*: in a given interaction context, the agent could observe the utterances generated by other agents. An assumption of appropriateness could be made, in which case hypotheses for the possible criteria weights that the agent utilized in the present scenario could be abduced. These hypotheses can be used by the agent itself as constraints that in turn govern its own utterance selection in similar social contexts.

[5] For example, what does it mean for utterance A to be equally less polite (e.g., 0.4) than utterance B as utterance B is less informative than utterance A?

Improved Operationalization of Criteria

Because our proposed framework relies on explicit operationalization of communicative criteria in order to rank candidate utterances, adapting and refining these operationalizations to new criteria, semantic representations, and NLG architectures will be an ongoing task. Adaptation will likely be fairly straightforward for criteria such as *correctness*, but other pragmatic and socio-linguistic criteria are more complex and leave room for future work. In particular, within DIARC the operationalizations of *politeness* and *brevity* can be improved and expanded. As alluded to earlier, brevity will require architectural integration with the lower-level NLG components such as the surface realizer and text-to-speech in order to calculate metrics for lexical and auditory brevity. This will be especially important when the spoken tempo of utterances can be manipulated (one can imagine a speed vs. intelligibility trade-off). Politeness is another criterion ripe for refinement. For example, though we modeled a scenario in which positive face (agent standing) was potentially threatened, a general framework to detect and evaluate threats to positive face is still needed (Briggs & Scheutz, 2014).

Conclusion

It is important that socially-embedded artificial agents generate speech in human-like ways in order for interaction with such agents to be truly natural. To this end, we have introduced and demonstrated a general method for modulating utterance selection based on an arbitrary number of social and pragmatic criteria. Our approach possesses an important set of novel features, including *extensibility* to additional sociolinguistic criteria, *adaptability* to changing situational context, and *agnosticism* with respect to underlying semantic representations. In a proof of concept demonstration, we showed how our approach can be integrated with a cognitive robotic architecture in order to generate flexible, socially-appropriate directives in a variety of contexts. Future work will be needed to extend the operationalization of the communicative criteria and devise mechanisms to learn the weights of the model through natural interaction. Overall, the present work moves us a step closer towards the goal of artificial agents that can communicate in the kinds of robust and socially-sensitive ways found in human language.

References

Bohnet, B., & Dale, R. (2005). Viewing referring expression generation as search. In *Proceedings of the 19th International Joint Conference on AI* (pp. 1004–1009).

Briggs, G., & Scheutz, M. (2013). A hybrid architectural approach to understanding and appropriately generating indirect speech acts. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence* (pp. 1213–1219).

Briggs, G., & Scheutz, M. (2014). Modeling blame to avoid positive face threats in natural language generation. In *Proceedings of the 8th International Conference on Natural Language Generation* (pp. 1–5). Philadelphia, PA.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge University Press.

Dale, R., & Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, *19*, 233–263.

Endrass, B., & André, E. (2014). Integration of cultural factors into the behavioural models of virtual characters. *Natural Language Generation in Interactive Systems*, 227–251.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*, 173–184.

Grice, P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics, 3: Speech Acts.*

Gupta, S., Walker, M. A., & Romano, D. M. (2007). How rude are you?: Evaluating politeness and affect in interaction. In *Affective Computing and Intelligent Interaction* (pp. 203–217). Springer.

Krahmer, E., & Van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, *38*, 173–218.

Mairesse, F., & Walker, M. A. (2011). Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*, *37*, 455–488.

Miller, C., Wu, P., & Funk, H. (2008). A computational approach to etiquette: Operationalizing brown and levinson's politeness model. *Intelligent Systems*, *23*, 28–35.

Moore, J. D., Porayska-Pomsta, K., Varges, S., & Zinn, C. (2004). Generating tutorial feedback with affect. In *Proceedings of the FLAIRS Conference* (pp. 923–928).

Nye, B. D., Graesser, A. C., & Hu, X. (2014). Autotutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, *24*, 427–469.

Schermerhorn, P. W., Kramer, J. F., Middendorff, C., & Scheutz, M. (2006). DIARC: A testbed for natural human-robot interaction. In *Proceedings of the AAAI Mobile Robot Workshop* (pp. 1972–1973).

Schulze, M. (2011). A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner election method. *Social Choice and Welfare*, *36*, 267–303.

Srinivasan, V., & Takayama, L. (2016). Help me please: Robot politeness strategies for soliciting help from humans. In *Proceedings of the 2016 Conference on Human Factors in Computing Systems* (pp. 4945–4955).

Strait, M., Canning, C., & Scheutz, M. (2014). Let me tell you! investigating the effects of robot communication strategies in advice-giving situations based on robot appearance, interaction modality and distance. In *Proceedings of the 9th International Conference on Human-Robot Interaction* (pp. 479–486).

Torrey, C., Fussell, S. R., & Kiesler, S. (2013). How a robot should give advice. In *Proceedings of the 8th International Conference on Human-Robot Interaction* (pp. 275–282).