Modeling Blame to Avoid Positive Face Threats in Natural Language Generation

Gordon Briggs
Human-Robot Interaction Laboratory
Tufts University
Medford, MA USA
gbriggs@cs.tufts.edu

Matthias Scheutz
Human-Robot Interaction Laboratory
Tufts University
Medford, MA USA
mscheutz@cs.tufts.edu

Abstract

Prior approaches to politeness modulation in natural language generation (NLG) often focus on manipulating factors such as the directness of requests that pertain to preserving the autonomy of the addressee (negative face threats), but do not have a systematic way of understanding potential impoliteness from inadvertently critical or blame-oriented communications (positive face threats). In this paper, we discuss ongoing work to integrate a computational model of blame to prevent inappropriate threats to positive face.

1 Introduction

When communicating with one another, people often modulate their language based on a variety of social factors. Enabling natural and human-like interactions with virtual and robotic agents may require engineering these agents to be able to demonstrate appropriate social behaviors. For instance, increasing attention is being paid to the effects of utilizing *politeness* strategies in both human-computer and human-robot dialogue interactions (Cassell and Bickmore, 2003; Torrey et al., 2013; Strait et al., 2014). This work has shown that, depending on context, the deployment of politeness strategies by artificial agents can increase human interactants' positive assessments of an agent along multiple dimensions (e.g. likeability).

However, while these studies investigated the human factors aspects of utilizing politeness strategies, they were not concerned with the natural language generation (NLG) mechanisms necessary to appropriately realize and deploy these strategies. Instead, there is a small, but growing, body of work on natural language generation architectures that seek to address this challenge (Gupta et al., 2007; Miller et al., 2008; Briggs and Scheutz, 2013). The common approach taken by these architectures is the operationalization of key factors in Brown and Levinson's seminal work on *politeness theory*, in particular, the degree to which an utterance can be considered a *face-threatening act* (FTA) (Brown and Levinson, 1987).

While this prior work demonstrates the abilities of these NLG architectures to successfully produce polite language, there remain some key challenges. Perhaps the most crucial question is: how does one calculate the degree to which an utterance is a FTA[1]? This is a complex issue, as not only is this value modulated by factors such as social distance, power, and context, but also the multifaceted nature of "face." An utterance may be polite in relation to *negative face* (i.e. the agent's autonomy), but may be quite impolite with regard to *positive face* (i.e. the agent's image and perceived character).

In this paper, we investigate the problem of modeling threats to positive face. First we discuss how prior work that has focused primarily on mitigating threats to negative face, and examine a specific example, taken from the human subject data of (Gupta et al., 2007), to show why accounting for positive face is necessary. Next, we discuss our proposed solution to begin to model threats to positive face– specifically, integrating a computational model of blame. Finally, we discuss the justification behind and limitations of this proposed approach.

2 Motivation

Brown and Levinson (1987) articulated a taxonomy of politeness strategies, distinguishing broadly between the notion of positive and negative politeness (with many distinct strategies for each). These categories of politeness correspond

[1] Less crucially, what is the appropriate notation for this value? It is denoted differently in each paper: Θ, W, and η.

to the concepts of positive and negative face, respectively. An example of a positive politeness strategy is the use of praise ("Great!"), whereas a common negative politeness strategy is the use of an *indirect speech act* (ISA), in particular, an indirect request. An example of an indirect request is the question, "Could you get me a coffee?", which avoids the autonomy-threatening direct imperative, while still potentially being construed as a request. This is an example of a conventionalized form, in which the implied request is more directly associated with the implicit form. Often considered even less of a threat to negative face are unconventionalized ISAs, which often require a deeper chain of inference to derive their implied meaning. It is primarily the modulation of the level of request indirectness that is the focus of (Gupta et al., 2007; Briggs and Scheutz, 2013).

To provide an empirical evaluation of their system, Gupta et al. (2007) asked human subjects to rate the politeness of generated requests on a five-point Likert scale in order of most rude (1) to to most polite (5). The results from (Gupta et al., 2007) for each of their politeness strategy categories are below:

1. Autonomy [3.4] (e.g. "Could you possibly do X for me?")
2. Approval [3.0] (e.g. "Could you please do X mate?")
3. Direct [2.0] (e.g. "Do X.")
4. Indirect [1.8] (e.g. "X is not done yet.")

This finding is, in some sense, counterintuitive, as unconventionalized request forms should be the least face-threatening. However, Gupta et al. (2007) briefly often an explanation, saying that the utterances generated in the indirect category sound a bit like a "complaint or sarcasm." We agree with this assessment. More precisely, while negative face is protected by the use of their unconventionalized ISAs, positive face was not.

To model whether or not utterances may be interpreted as being complaints or criticisms, we seek to determine whether or not they can be interpreted as an act of *blame*[2].

[2]What the precise ontological relationship is between concepts such as complaining, criticizing, and blaming is beyond the scope of this paper.

3 Approach

Like praise, blame (its negative counterpart) is both a cognitive and social phenomenon (Malle et al., 2012). The cognitive component pertains to the internal attitudes of an agent regarding another agent and their actions, while the social component involves the expression of these internal attitudes through communicative acts. To achieve blame-sensitivity in NLG, we need to model both these aspects. In the following sections, we briefly discuss how this could be accomplished.

3.1 Pragmatic and Belief Reasoning

Before a speaker S can determine the high-level perlocutionary effects of an utterance on an addressee (H) vis-á-vis whether or not they feel criticized or blamed, it is first necessary to determine the precise set of beliefs and intentions of the addressee upon hearing an utterance u in context c. We denote this updated set of beliefs and intentions $\Psi_H(u, c)$. Note that this set is a *model* of agent H's beliefs and intentions from the speaker S's perspective, and not necessarily equivalent to the actual belief state of agent H. In order to perform this mental modeling, we utilize a reasoning system similar to that in (Briggs and Scheutz, 2011). This pragmatic reasoning architecture utilizes a set of rules of the form:

$$[[U]]_C := \phi_1 \wedge ... \wedge \phi_n$$

where U denotes an utterance form, C denotes a set of contextual constraints that must hold, and ϕ denotes a belief update predicate. An utterance form is specified by $u = UtteranceType(\alpha, \beta, X, M)$, where $UtteranceType$ denotes the dialogue turn type (e.g. statement, y/n-question), α denotes the speaker of the utterance u, β denotes the addressee of the utterance, X denotes the surface semantics of the utterance, and M denotes a set of sentential modifiers. An example of such a pragmatic rule is found below:

$$[[Stmt(S, H, X, \{\})]]_\emptyset := want(S, bel(H, X))$$

which denotes that a statement by the speaker S to an addressee H that X holds should indicate that, "S wants H to believe X," in all contexts (given the empty set of contextual constraints). If this rule matches a recognized utterance (and the contextual constraints are satis-

fied, which is trivial in this case), then the mental model of the addressee is updated such that: $want(S, bel(H, X)) \in \Psi_H(u, c)$.

Of particular interest with regard to the Gupta et al. (2007) results, Briggs and Scheutz (2011) describe how they can use their system to understand the semantics of the adverbial modifier "yet," which they describe as being indicative of mutually understood intentionality. More accurately, "yet," is likely indicative of a belief regarding *expectation* of an action being performed or state being achieved. Therefore, a plausible pragmatic rule to interpret, "X is not done yet," could be:

$$[[Stmt(S, H, \neg done(X), \{yet\})]]_\emptyset := \\ want(S, bel(H, \neg done(X))) \wedge \\ expects(S, done(X))$$

Furthermore, in a cooperative, task-driven context, such as that described in (Gupta et al., 2007), it would not be surprising for an interactant to infer that this expectation is further indicative of a belief in a particular intention or a task-based obligation to achieve X.[3]

As such, if we consider an utterance u_d as being a standard direct request form (strategy 3), and an utterance u_y as being an indirect construction with a yet modifier (strategy 4), the following facts may hold:

$$bel(S, promised(H, S, X, t_p)) \notin \Psi_H(u_d, c)$$

$$bel(S, promised(H, S, X, t_p)) \in \Psi_H(u_y, c)$$

If S is making a request to H, there is no believed agreement to achieve X. However, if "yet," is utilized, this may indicate to H a belief that S thinks there is such an agreement.

Having calculated an updated mental model of the addressee's beliefs after hearing a candidate utterance u, we now can attempt to infer the degree to which u is interpreted as an act of criticism or blame.

3.2 Blame Modeling

Attributions of blame are influenced by several factors including, but not limited to, beliefs about an agent's intentionality, capacity, foreknowledge, obligations, and possible justifications (Malle et al., 2012). Given the centrality of intentionality in blame attribution, it is unsurprising that current computational models involve reasoning within a symbolic BDI (belief, desire, intention) framework, utilizing rules to infer an ordinal degree of blame based on the precise set of facts regarding these factors (Mao and Gratch, 2012; Tomai and Forbus, 2007). A rule that is similar to those found in these systems is:

$$bel(S, promised(H, S, X, t_p)) \wedge bel(S, \neg X) \wedge \\ bel(S, (t > t_p)) \wedge bel(S, capable\_of(H, X)) \\ \Rightarrow blames(S, H, high)$$

that is to say, if agent S believes agent H promised to him or her to achieve X by time t_p, and S believes X has not been achieved and the current time t is past t_p, and S believes H is capable of fulfilling this promise, then S will blame H to a high degree. Continuing our discussion regarding the perlocutionary effects of u_d and u_y, it is likely then that: $blames(S, H, high) \notin \Psi_H(u_d, c)$ and $blames(S, H, high) \in \Psi_H(u_y, c)$.

3.3 FTA Modeling

Having determined whether or not an addressee would feel criticized or blamed by a particular candidate utterance, it is then necessary to translate this assessment back into the terms of FTA-degree (the currency of the NLG system). This requires a function $\beta(\Psi)$ that maps the ordinal blame assessment of the speaker toward the hearer based on a set of beliefs Ψ, described in the previous section, to a numerical value than can be utilized to calculate the severity of the FTA (e.g. $blames(S, H, high) = 9.0$, $blames(S, H, medium) = 4.5$). For the purposes of this paper we adopt the theta-notation of Gupta et al. (2007) to denote the degree to which an utterance is a FTA. With the β function, we can then express the blame-related FTA severity of an utterance as:

$$\Theta_{blame}(u, c) = \beta_H(\Psi_H(u, c)) - \alpha(c) \cdot \beta_S(\Psi_S)$$

where β_H denotes the level of blame the speaker believes the hearer has inferred based on the addressee's belief state after hearing utterance u with context c ($\Psi_H(u, c)$)). β_S denotes the level of blame the speaker believes is appropriate given his or her current belief state. Finally, $\alpha(c)$ denotes a

[3] How precisely this reasoning is and/or ought to be performed is an important question, but is outside the scope of this paper.

multiplicative factor that models the appropriateness of blame given the current social context. For instance, independent of the objective blameworthiness of a superior, it may be inappropriate for a subordinate to criticize his or her superior in certain contexts.

Finally, then, the degree to which an utterance is a FTA is the sum of all the contributions of evaluations of possible threats to positive face and possible threats to negative face:

$$\Theta(u,c) = \sum_{p \in P} \Theta_p(u,c) + \sum_{n \in N} \Theta_n(u,c)$$

where P denotes the set of all possible threats to positive face (e.g. blame) and N denotes the set of all possible threats to negative face (e.g. directness).

We can see how this would account for the human-subject results from (Gupta et al., 2007), as conventionally indirect requests (strategies 1 and 2) would not produce large threat-value contributions from either the positive or negative FTA components. Direct requests (strategy 3) would, however, potentially produce a large Θ_N contribution, while their set of indirect requests (strategy 4) would trigger a large Θ_P contribution.

4 Discussion

Having presented an approach to avoid certain types of positive-FTAs through reasoning about blame, one may be inclined to ask some questions regarding the justification behind this approach. Why should we want to better model one highly complex social phenomenon (politeness) through the inclusion of a model of another highly complex social phenomenon (blame)? Does the integration of a computational model of blame actually add anything that would justify the effort?

At a superficial level, it does not. The criticism/blame-related threat of a specific speech act can be implicitly factored into the base FTA-degree evaluation function supplied to the system, determined by empirical data or designer-consensus as is the case of (Miller et al., 2008). However, this approach is limited in a couple ways. First, this does not account for the fact that, in addition to the set of social factors Brown and Levinson articulated, the appropriateness of an act of criticism or blame is also dependent on whether or not it is *justified*. Reasoning about whether or not an act of blame is justified requires: a computational model of blame.

Second, the inclusion of blame-reasoning within the larger scope of the entire agent architecture may enable useful behaviors both inside and outside the natural language system. There is a growing community of researchers interested in developing ethical-reasoning capabilities for autonomous agents (Wallach and Allen, 2008), and the ability to reason about blame has been proposed as one key competency for such an ethically-sensitive agent (Bello and Bringsjord, 2013). Not only is there interest in utilizing such mechanisms to influence general action-selection in autonomous agents, but there is also interest in the ability to understand and generate valid explanations and justifications for adopted courses of action in ethically-charged scenarios, which is of direct relevance to the design of NLG architectures.

While our proposed solution tackles threats to positive face that arise due to unduly critical/blame-oriented utterances, there are many different ways of threatening positive face aside from criticism/blame. These include phenomena such as the discussion of inappropriate/sensitive topics or non-cooperative behavior (e.g. purposefully ignoring an interlocutor's dialogue contribution). Indeed, empirical results show that referring to an interlocutor in a dyadic interaction using an impersonal pronoun (e.g. "someone") may constitute another such positive face threat (De Jong et al., 2008). Future work will need to be done to develop mechanisms to address these other possible threats to positive face.

5 Conclusion

Enabling politeness in NLG is a challenging problem that requires the modeling of a host of complex, social psychological factors. In this paper, we discuss ongoing work to integrate a computational model of blame to prevent inappropriate threats to positive face that can account for prior human-subject data. As an ongoing project, future work is needed to further test and evaluate this proposed approach.

Acknowledgments

We would like to thank the reviewers for their helpful feedback. This work was supported by NSF grant #111323.

References

Paul Bello and Selmer Bringsjord. 2013. On how to build a moral machine. *Topoi*, 32(2):251–266.

Gordon Briggs and Matthias Scheutz. 2011. Facilitating mental modeling in collaborative human-robot interaction through adverbial cues. In *Proceedings of the SIGDIAL 2011 Conference*, pages 239–247, Portland, Oregon, June. Association for Computational Linguistics.

Gordon Briggs and Matthias Scheutz. 2013. A hybrid architectural approach to understanding and appropriately generating indirect speech acts. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some universals in language usage*. Cambridge University Press.

Justine Cassell and Timothy Bickmore. 2003. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction*, 13(1-2):89–132.

Markus De Jong, Mariët Theune, and Dennis Hofs. 2008. Politeness and alignment in dialogues with a virtual guide. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, pages 207–214. International Foundation for Autonomous Agents and Multiagent Systems.

Swati Gupta, Marilyn A Walker, and Daniela M Romano. 2007. How rude are you?: Evaluating politeness and affect in interaction. In *Affective Computing and Intelligent Interaction*, pages 203–217. Springer.

Bertram F Malle, Steve Guglielmo, and Andrew E Monroe. 2012. Moral, cognitive, and social: The nature of blame. *Social thinking and interpersonal behavior*, 14:313.

Wenji Mao and Jonathan Gratch. 2012. Modeling social causality and responsibility judgment in multiagent interactions. *Journal of Artificial Intelligence Research*, 44(1):223–273.

Christopher A Miller, Peggy Wu, and Harry B Funk. 2008. A computational approach to etiquette: Operationalizing brown and levinson's politeness model. *Intelligent Systems, IEEE*, 23(4):28–35.

Megan Strait, Cody Canning, and Matthias Scheutz. 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 2014 ACM/IEEE international conference on Human-robot interaction*, pages 479–486. ACM.

Emmett Tomai and Ken Forbus. 2007. Plenty of blame to go around: a qualitative approach to attribution of moral responsibility. Technical report, DTIC Document.

Cristen Torrey, Susan R Fussell, and Sara Kiesler. 2013. How a robot should give advice. In *Human-Robot Interaction (HRI), 2013 8th ACM/IEEE International Conference on*, pages 275–282. IEEE.

Wendell Wallach and Colin Allen. 2008. *Moral machines: Teaching robots right from wrong*. Oxford University Press.