Facilitating the Communication of Politeness through Fine-Grained Paraphrasing

Liye Fu
Cornell University
liye@cs.cornell.edu

Susan R. Fussell
Cornell University
sfussell@cornell.edu

Cristian Danescu-Niculescu-Mizil
Cornell University
cristian@cs.cornell.edu

Abstract

Aided by technology, people are increasingly able to communicate across geographical, cultural, and language barriers. This ability also results in new challenges, as interlocutors need to adapt their communication approaches to increasingly diverse circumstances. In this work, we take the first steps towards automatically assisting people in adjusting their language to a specific communication circumstance.

As a case study, we focus on facilitating the accurate transmission of pragmatic intentions and introduce a methodology for suggesting paraphrases that achieve the intended level of politeness under a given communication circumstance. We demonstrate the feasibility of this approach by evaluating our method in two realistic communication scenarios and show that it can reduce the potential for misalignment between the speaker's intentions and the listener's perceptions in both cases.

1 Introduction

Technological developments have greatly enhanced our communication experience, providing the opportunity to overcome geographic, cultural and language barriers to interact with people from different backgrounds in diverse settings (Herring, 1996). However, this opportunity brings additional challenges for the interlocutors, as they need to adjust their language to increasingly diverse communication circumstances.

As humans, we often make conscious attempts to account for the communication setting. For instance, we may simplify our expressions if we know our listener has relatively limited language proficiency, and we tend to be more polite towards people with higher status. However, managing these stylistic adjustments can be cognitively taxing, especially when we are missing relevant information—e.g., the language proficiency or the status of a conversational partner we meet online.

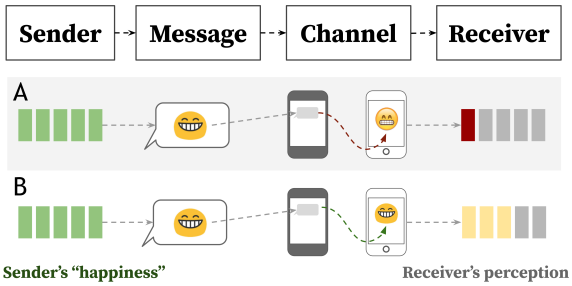

Figure 1: Berlo's Sender-Message-Channel-Receiver model suggests that the intended and perceived style of a message can be misaligned if: **A**. the `channel` does not faithfully transmit the message, or **B**. the `receiver` has a different reading of the message compared to the `sender`. Examples are inspired by Miller et al. (2016).

If we do not adjust our language, we risk not properly conveying our pragmatic intentions (Thomas, 1983). In particular, Berlo's Sender-Message-Channel-Receiver model (Berlo, 1960) points to two potential circumstance-specific causes for misalignments between intentions and perceptions (Figure 1). In this work we explore a method for assisting speakers to avoid such misalignments by suggesting for each message a paraphrase that is more likely to convey the original pragmatic intention when communicated in a given circumstance, as determined by the properties of the `sender`, `channel`, and `receiver`.

As a case study, in this work, we focus on one particular pragmatic aspect: politeness. It is important to assist people to accurately transmit their intended politeness, as this interpersonal style (Biber, 1988) plays a key role in social interactions (Burke and Kraut, 2008; Murphy, 2014; Hu et al., 2019; Maaravi et al., 2019). Furthermore, politeness is known to be a circumstance-sensitive phenomenon (Kasper, 1990; Herring, 1994; Forgas, 1998; Mousavi and Samar, 2013), making it a good case for our study. Concretely, we propose the task of generating a paraphrase for a given

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 5127–5140,
November 16–20, 2020.

message that is more likely to deliver the intended level of politeness after transmission (henceforth *intention-preserving*), considering the properties of the `sender`, `channel`, and `receiver` (Section 3).

Taking the properties of the `channel` into account is important because communication `channels` may not always faithfully deliver messages (Figure 1A). For example, in translated communication, politeness signals can often be lost or corrupted (Allison and Hardin, 2020). To demonstrate the potential of our framework in mitigating `channel`-induced misunderstandings, we apply it to suggest paraphrases that are *safer to transmit*—i.e., less likely to have their politeness altered—over a commercial machine translation service.

We also need to account for the fact that the `sender` and `receiver` can have different interpretations of the same message (Figure 1B). For example, people may perceive politeness cues differently depending on their cultural background (Thomas, 1983; Riley, 1984). In our second application scenario, the interlocutors' perceptions of politeness are misaligned, and we aim to suggest paraphrases that reduce the potential for misinterpretation.

To successfully produce such circumstance-sensitive paraphrases, we need to depart from existing style transfer methodology (see Li et al., 2020 for a survey, and Madaan et al., 2020 for politeness transfer in particular). First, since we must account for arbitrary levels of misalignment, we need *fine-grained* control over the target stylistic level, as opposed to binary switches (e.g., from impolite to polite). Second, we need to determine the target stylistic level at run time, in an *ad hoc* fashion, rather than assuming predefined targets.

To overcome these new technical challenges, we start from the intuition that the same level of politeness can be conveyed through different combinations of *pragmatic strategies* (Lakoff, 1973; Brown and Levinson, 1987), with some being more appropriate to the given circumstance than others. We consider a classic two-step approach (Section 4), separating *planning*—choosing a viable combination of strategies that can achieve a desired stylistic level in a particular circumstance—, from the step of *realization*—incorporating this plan into generation outputs. For a given fine-grained target stylistic level (i.e., the level intended by the sender), we find the optimal strategy plan via Integer Linear Programming (ILP). We then realize this plan using a modification of the 'Delete-Retrieve-Generate' (DRG) paradigm (Li et al., 2018) that allows for strategy-level control in generation.

Our experimental results indicate that in both our application scenarios, our method can suggest paraphrases that narrow the potential gap between the intended and perceived politeness, and thus better preserve the sender's intentions. These results show that automated systems have the potential to help people better convey their intentions in new communication circumstances, and encourage further work exploring the feasibility and implications of such communication assistance applications.

To summarize, in this work, we motivate and formulate the task of circumstance-sensitive intention-preserving paraphrasing (Section 3). Focusing on the case of pragmatic intentions, we introduce a model for paraphrasing with fine-grained politeness control (Section 4). We evaluate our method in two realistic communication scenarios to demonstrate the feasibility of the approach (Section 5).

2 Further Related Work

Style transfer. There has been a wide range of efforts in using NLP techniques to generate alternative expressions, leading to tasks such as text simplification (see Shardlow, 2014 for a survey), or more generally, paraphrase generation (Meteer and Shaked, 1988; Quirk et al., 2004; Fu et al., 2019, inter alia). When such paraphrasing effort is focused on the stylistic aspect, it is also referred to as text style transfer, which has attracted a lot of attention in recent years (Xu et al., 2012; Ficler and Goldberg, 2017; Fu et al., 2018; Prabhumoye et al., 2018, inter alia). While these tasks are focused on satisfying specific predefined linguistic properties at the utterance-level, they are not designed for fine-grained adjustments to changing non-textual communication circumstances.

Controllable generation. Style transfer or paraphrasing can both be seen as a special case of the broader task of *controllable* text generation (Hu et al., 2017; Keskar et al., 2019; Dathathri et al., 2020, inter alia). While not focused on paraphrasing, relevant work in this area aims at controling the level of politeness for translation (Sennrich et al., 2016) or dialog response (Niu and Bansal, 2018).

AI-assisted communications or writing. Beyond paraphrasing, AI tools have been used to provide communication or writing assistance in diverse settings: from the mundane task of grammar and spell checking (Napoles et al., 2019; Stevens, 2019), to

creative writing (Clark et al., 2018), to negotiations (Zhou et al., 2019), and has led to discussions of ethical implications (Hancock et al., 2020).

Models of communication. While Berlo's model provides the right level of abstraction for inspiring our application scenarios, many other models exist (Velentzas and Broni, 2014; Barnlund, 2017), most of which are under the influence of the Shannon–Weaver model (Shannon and Weaver, 1963).

3 Task Formulation

Given a message that a `sender` attempts to communicate to a `receiver` over a particular communication `channel`, the task of circumstance-sensitive intention-preserving paraphrasing is to generate a paraphrase that is more likely to convey the intention of the `sender` to the `receiver` after transmission, under the given communication circumstance.

Formulation. To make this task more tractable, our formulation considers a single gradable stylistic aspect of the message that can be realized through a collection of *pragmatic strategies* (denoted as $\mathcal{S}$). While in this work we focus on politeness, other gradable stylistic aspects might include formality, humor and certainty.

We can then formalize the relevant features of the communication circumstance as follows:

1. For the communication `channel`, we consider whether it can safely transmit each strategy $s \in \mathcal{S}$. In particular, $f_c\,(s) = 1$ indicates that strategy s is *safe* to use, whereas $f_c\,(s) = 0$ implies that s is *at-risk* of being lost.

2. For the `sender` and `receiver`, we quantify the level of the stylistic aspect each of them perceive in a combination of pragmatic strategies via two mappings $f_{send} : \mathcal{P}(\mathcal{S}) \rightarrow \mathbb{R}$ and $f_{rec} : \mathcal{P}(\mathcal{S}) \rightarrow \mathbb{R}$, respectively, with $\mathcal{P}(\mathcal{S})$ denoting the powerset of $\mathcal{S}$.

With our focus on politeness, our task can then be more precisely stated as follows: given an input message m, we aim to generate a politeness paraphrase for m, under the circumstance specified by (f_{send}, f_c, f_{rec}), such that the level of politeness perceived by the `receiver` is as close to that intended by the `sender` as possible.

We show that our theoretically-grounded formulation can model naturally-occurring challenges in communication, by considering two possible application scenarios, each corresponding to a source of misalignment highlighted in Figure 1.

Application A: translated communication. We first consider the case of conversations mediated by translation services, where `channel`-induced misunderstandings can occur (Figure 1A): MT models may systematically drop certain politeness cues due to technical limitations or mismatches between the source and target languages.

For instance, despite the difference in intended politeness level (indicated in parentheses) of the following two versions of the same request,[1]

> Could you please proofread this article? (POLITE)
> Can you proofread this article? (SOMEWHAT POLITE)

Microsoft Bing Translator would translate both versions to the same utterance in Chinese.[2] By dropping the politeness marker 'please', and not making any distinction between 'could you' and 'can you', the message presented to the Chinese receiver is likely to be more imposing than originally desired by the English sender.

To avoid such `channel`-induced misunderstandings, the sender may consider using only strategies that are known to be safe with the specific MT system they use.[3] However, since the inner mechanics of such systems are often opaque (and in constant flux), the sender would benefit from automatic guidance in constructing such paraphrases.

Application B: misaligned perceptions. We then consider the case when `senders` and `receivers` with differing perceptions interact. Human perceptions of pragmatic devices are subjective, and it is not uncommon to observe different interpretations of the same utterance, or pragmatic cues within, leading to misunderstandings (Thomas, 1983; Kasper, 1990) (Figure 1B). For instance, a study comparing Japanese speakers' and American native English speakers' perceptions of English requests find that while the latter group takes the request '*May I borrow a pen?*' as strongly polite, their Japanese counterparts regard the expression as almost neutral (Matsuura, 1998). In this case, if a native speaker still wishes to convey their good will, they need to find a paraphrase that would be perceived as strongly polite by Japanese speakers.

[1] Annotations from 5 native speakers on a 7-point Likert scale ranging from VERY IMPOLITE to VERY POLITE.

[2] Translating on May, 2020 to 你能校对这篇文章吗?

[3] E.g., they might consider expressing gratitude (e.g., 'thanks!') rather than relying on subjunctive ('could you').

4 Method

When compared to style transfer tasks, our circumstance-sensitive intention-preserving paraphrasing task gives rise to important new technical challenges. First, in order to minimize the gap in perceptions, we need to have fine-grained control over the stylistic aspect, as opposed to switching between two pre-defined binarized targets (e.g., polite vs. impolite). Second, the desired degree of change is only determined at run-time, depending on the speaker's intention and on the communication circumstance. We address these challenges by developing a method that allows for *ad hoc* and *fine-grained* paraphrase planning and realization.

Our solution starts from a strategy-centered view: instead of aiming for monolithic *style labels*, we think of *pragmatic strategies* as (stylistic) LEGO bricks. These can be stacked together in various combinations to achieve similar stylistic levels. Depending on the circumstance, some bricks might, or might not, be available. Therefore, given a message with an intended stylistic level, our goal is to find the optimal collection of available bricks that can convey the same level—ad hoc fine-grained planning. Given this optimal collection, we need to assemble it with the rest of the message into a valid paraphrase—fine-grained realization.

Politeness strategies. In the case of politeness, we derive the set of pragmatic strategies from prior work (Danescu-Niculescu-Mizil et al., 2013; Voigt et al., 2017; Yeomans et al., 2019). We focus on strategies that are realized through local linguistic *markers*. For instance, the Subjunctive strategy can be realized through the use of markers like *could you* or *would you*. In line with prior work, we further assume that markers realizing the same strategy has comparable *strength* in exhibiting politeness and are subject to the same constraints. The full list of 18 strategies we consider (along with their example usages) can be found in Table 1. Strategy extraction code is available in ConvoKit.[4]

Ad hoc fine-grained planning. Our goal is to find a target strategy combination that is estimated to provide a comparable pragmatic force to the sender's intention, using only strategies appropriate in the current circumstance. To this end, we devise an Integer Linear Programming (ILP) formulation that can efficiently search for the desired strategy combination to use (Section 4.1).

[4] https://convokit.cornell.edu.

Strategy	Example usage
Actually	it **actually** needs to be ...
Adverb.Just	i **just** noticed that ...
Affirmation	**excellent point**, i have added it ...
Apology	**sorry** to be off-topic, but ...
By.The.Way	okay - **by the way**, do you want me ...?
Conj.Start	**so** where is the article ?
Filler	**uh**, hey, can you...?
For.Me	is it alright **for me** to archive it now?
For.You	i can fetch one **for you** in a moment! ...
Gratitude	**thanks** for the info , ...
Greeting	**hey** simon , help is needed if possible ...
Hedges	**maybe** some kind of citation is needed ...
Indicative	**can you** create one for me?
Please	can you **please** check it?
Please.Start	**please** stop . if you continue ...
Reassurance	**no problem**, happy editing. ...
Subjunctive	..., **could you** check?
Swearing	what **the heck** are you talking about?

Table 1: Politeness strategies we consider, along with example usage and example markers (in bold). More details for the strategies can be found in Table A1 in the Appendix.

Fine-grained realization. To train a model that learns to merge the strategy plan into the original message in the absence of parallel data, we take inspirations from the DRG paradigm (Li et al., 2018), originally proposed for style transfer tasks. We adapt this paradigm to allow for direct integration with strategy-level planning, providing finer-grained control over realization (Section 4.2).

4.1 Fine-Grained Strategy Planning

Formally, given a message m using a set of strategies $\mathcal{S}_{in}$, under a circumstance specified by (f_{send}, f_c, f_{rec}), the planning goal is to find the set of strategies $\mathcal{S}_{out} \subseteq \mathcal{S}$ such that $f_c\ (s) = 1, \forall s \in \mathcal{S}_{out}$ —i.e., they can be safely transmitted through the communication `channel`—and $f_{send}\ (\mathcal{S}_{in}) \approx f_{rec}\ (\mathcal{S}_{out})$—i.e., the resultant `receiver` perception is similar to the intention the `sender` had when crafting the original message.

Throughout, we assume that both perception mappings f_{send} and f_{rec} are linear functions:

$$f_{send}(\mathcal{S}_{in}) = \sum_{s \in \mathcal{S}} a_s \mathbb{1}_{\mathcal{S}_{in}}(s) + a_0$$

$$f_{rec}(\mathcal{S}_{out}) = \sum_{s \in \mathcal{S}} b_s \mathbb{1}_{\mathcal{S}_{out}}(s) + b_0$$

where the linear coefficients a_s and b_s are reflective of the strength of a strategy, as perceived by the `sender` and `receiver`, respectively.[5]

[5] We acknowledge that considering only linear models may result in sub-optimal estimations.

Naive approach. One greedy type of approach to this problem is to consider each at-risk strategy $s \in \mathcal{S}_{in}$ at a time, and replace s with a safe strategy s' that is closest in strength. Mathematically, this can be written as $s' = \arg\min_{\hat{s} \in \mathcal{S}, f_c(\hat{s})=1} |a_s - b_{\hat{s}}|$. In our analogy, this amounts to reconstructing a LEGO model by replacing each 'lost' brick with the most similar brick that is available.

Our approach: ILP formulation. The greedy approach, while easy to implement, can not consider solutions that involve an alternative *combination* of strategies. In order to more thoroughly search for an appropriate strategy plan in the space of possible solutions in a flexible and efficient manner, we translate this problem into an ILP formulation.[6]

Our objective is to find a set of safe strategies $\mathcal{S}_{out}$ that will be perceived by the `receiver` as close as possible to the `sender`'s intention, i.e., one that that minimizes $|f_{send}(\mathcal{S}_{in}) - f_{rec}(\mathcal{S}_{out})|$.

To this end, we introduce a binary variable x_s for each strategy s in $\mathcal{S}$, where we take $x_s = 1$ to mean that strategy s should be selected to be present in the suggested alternative strategy combination $\mathcal{S}_{out}$. We can identify the optimal value of x_s (and thus the optimal strategy set $\mathcal{S}_{out}$) by solving the following ILP problem:[7]

$$\begin{aligned} \text{MIN} \quad & y \\ \text{subj to} \quad & (\textstyle\sum a_s \mathbb{1}_{\mathcal{S}_{in}}(s) + a_0) - (\sum b_s x_s + b_0) \leq y \\ & (\textstyle\sum b_s x_s + b_0) - (\sum a_s \mathbb{1}_{\mathcal{S}_{in}}(s) + a_0) \leq y \\ & x_s \leq f_c(s), x_s \in \{0,1\}, \forall s \in \mathcal{S} \end{aligned}$$

which is a rewriting our objective to minimize $|f_{send}(\mathcal{S}_{in}) - f_{rec}(\mathcal{S}_{out})|$ that satisfies the linearity requirement of ILP via an auxiliary variable y, and where our target variables x_s replace the indicator function $\mathbb{1}_{\mathcal{S}_{out}}(s)$ in the linear expression of f_{rec}.

The `channel` constraints are encoded by the additional constraints $x_s \leq f_c(s)$, allowing only safe strategies (i.e., those for which $f_c(s) = 1$) to be included. Additional strategy-level constraints can be similarly specified through this mechanism to obtain strategy plans that are easier to realize in natural language (Section C in the Appendix).

4.2 Fine-Grained Realization

To transform the ILP solutions into natural language paraphrases, we build on the general DRG framework, which has shown strong performance in style transfer without parallel data.[8] We modify this framework to allow for the fine-grained control needed to realize strategy plans.

As the name suggests, the vanilla DRG framework consists of three steps. With *delete*, lexical markers (n-grams) that are strongly indicative of style are removed, resulting in a 'style-less' intermediate text. In the *retrieve* step, target markers are obtained by considering those used in training examples that are similar to the input but exhibit the desired property (e.g., target sentiment valence). Finally, in the *generate* step, the generation model merges the desired target markers with the style-less intermediate text to create the final output.

Importantly, the DRG framework is primarily designed to select to-be-inserted markers based on pre-defined binary style classes. As such, it cannot directly allow the ad hoc fine-grained control needed by our application. We now explain our modifications in detail (follow the sketch of our pipeline in Figure 2):

Plan (instead of Retrieve). We first perform a Plan step, which substitutes the Retrieve step in DRG, but it is performed first in our pipeline as our version of the Delete step is dependent on the planning results. For an input message, we identify the politeness strategies it contains and set up the corresponding ILP problem (Section 4.1) to obtain their *functional alternatives*. By factoring in the communication circumstance into the ILP formulation, we obtain an ad hoc strategy plan to achieve the intended level of politeness. This is in contrast with the Retrieve step in DRG, in which target markers from similar-looking texts are used for direct *lexical substitution*.

Delete. Instead of identifying style-bearing lexical markers to delete with either frequency-based heuristics (Li et al., 2018), or sentence context (Sudhakar et al., 2019), we rely on linguistically informed politeness strategies. To prepare the input message for the new strategy plan, we compare the strategy combination from the ILP solution with those originally used. We then *selectively* remove strategies that do not appear in the ILP solution by deleting the corresponding markers found in the input message. As such, in contrast with DRG, our post-deletion context is not necessarily style-less, and it is also possible that no deletion is performed.

[6] A brute force alternative would inevitably be less scalable.

[7] All summations are over the entire strategy set $\mathcal{S}$. Throughout, we use the `PuLP` package (Mitchell et al., 2011) with `GLPK` solver to obtain solutions.

[8] Since the politeness strategies we consider are local, they fit the assumptions of DRG framework well.

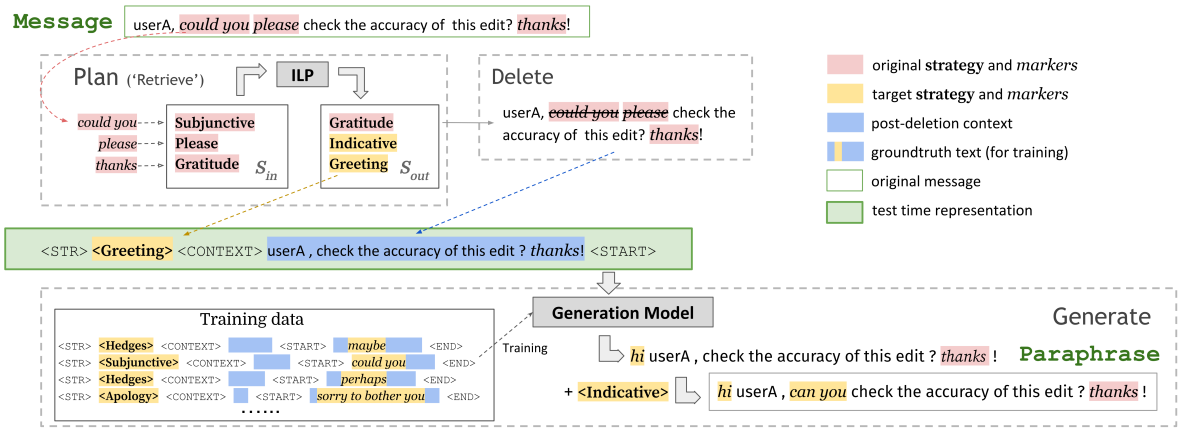

Figure 2: Sketch of our pipeline for generating politeness paraphrases. Given an input message, we first identify the politeness strategies ($\mathcal{S}_{in}$) and the corresponding markers it contains. In the **plan** step, we use ILP to compute a target strategy combination ($\mathcal{S}_{out}$) that is appropriate under the circumstance. We then **delete** markers corresponding to strategies that need to be removed to obtain the post-deletion context. Finally, we sequentially insert the new strategies from the ILP solution into this context to **generate** the final output.

Generate. Finally, we need to generate fluent utterances that integrate the strategies identified by the Plan step into the post-deletion context. To this end, we adapt G-GST (Sudhakar et al., 2019), whose generation model is fine-tuned to learn to integrate lexical markers into post-deletion context. To allow smooth integration of the ILP solution, we instead train the generation model to incorporate politeness strategies directly.

Concretely, training data exemplifies how each target strategy can be integrated into various post-deletion contexts. This data is constructed by finding GROUNDTRUTH utterances containing markers corresponding to a certain STRATEGY, and removing them to obtain the post-deletion CONTEXT. These training instances are represented as (STRATEGY, CONTEXT, GROUNDTRUTH) tuples separated by special tokens (examples in Figure 2). The model is trained to minimize the reconstruction loss.[9]

At test time, we sequentially use the model to integrate each STRATEGY from the plan into the post-deletion CONTEXT. We perform beam search of size 3 for each strategy we attempt to insert and select the output that best matches the intended level of politeness as the paraphrase suggestion.[10]

5 Evaluation

To test the feasibility of our approach, we set up two parallel experiments with different circumstance specifications, so that each illustrates one potential source of misalignment as described in Section 3.[11]

5.1 Experiments

Data. We use the annotations from the Wikipedia section of the Stanford Politeness Corpus (henceforth annotations) to train perception models that will serve as approximations of f_{send} and f_{rec}. In this corpus, each utterance was rated by 5 annotators on a 25-point scale from very impolite to very polite, which we rescale to the $[-3, 3]$ range.

To train the generation model, we randomly sample another (unannotated) collection of talk-page messages from WikiConv (Hua et al., 2018). For each strategy, we use 1,500 disjoint instances for training (27,000 in total, 2000 used for validation) and additionally resource 200 instances per strategy as test data. Both the Stanford Politeness Corpus and WikiConv are retrieved from ConvoKit (Chang et al., 2020b).

Experiment A: translated communication. We first consider MT-mediated English to Chinese communication using Microsoft Translator, where `channel`-induced misunderstandings may occur.

For this specific `channel`, we estimate its f_c by performing back-translation[12] (Tyupa, 2011) on a

[9] We adapted the implementation from Sudhakar et al. (2019) to incorporate our modification described above, and we use their default training setup.

[10] We set an upper bound of at most 3 new strategies to be introduced to keep sequential insertion computationally manageable. This is a reasonable assumption for short utterances.

[11] Code and data is available at `https://github.com/CornellNLP/politeness-paraphrase`.

[12] Back-translation refers to the process of translating the

sampled set of utterances from the collection of Stack Exchange requests from the Stanford Politeness Corpus. We consider a strategy s to be at-risk under this MT-mediated `channel` if the majority of messages using s have back-translations that no longer uses it. We identify four at-risk strategies, leading to the following channel specification: $f_c(s) = 0$, if $s \in \{\text{Subjunctive, Please, Filler, Swearing}\}$; $f_c(s) = 1$ otherwise.

For the `sender` and the `receiver`, we make the simplifying assumption that they both perceive politeness similar to a prototypical 'average person' (an assumption we address in the next experiment), and take the average scores from the annotations to train a linear regression model f_{avg} to represent the perception model, i.e., $f_{send} = f_{rec} = f_{avg}$.

We retrieve test data corresponding to the four at-risk strategy types as test messages (4×200 in total). We estimate the default perception gap (i.e., when **no intervention** takes place) by comparing the intended level of politeness in the original message and the level of politeness of its back-translation, which roughly approximates what the `receiver` sees after translation, following Tyupa (2011). This way, we can avoid having to compare politeness levels across different languages.

Experiment B: misaligned perceptions. We then consider communication between individuals with misaligned politeness perceptions. Under this circumstance, we assume a perfect `channel`, which allows any strategy to be safely transmitted, i.e., $f_c(s) = 1, \forall s \in S$. We then consider the top 5 most prolific annotators as potential `senders` and `receivers`. To obtain f_{send} (and f_{rec}), we use the respective annotator's annotations to train an individual linear regression model.[13]

We take all permutations of `(sender, receiver)` among the chosen annotators, resulting in 20 different directed pairs. For each pair, we select as test data the top 100 utterances with the greatest (expected) perception gap in the test set. We take the default perception gap within the pair (with **no intervention**) as the difference between the `sender`'s intended level of politeness (as judged by f_{send}) and the `receiver`'s perceived level of politeness (as judged by f_{rec}).

translated text back into the source language.

[13]Details about the choice of annotators and their perception models are described in Section B in the Appendix. While in practice individual perception models may not be available, they could potentially be approximated based on annotations from people with similar (cultural) backgrounds.

Baselines. Beyond the base case with no intervention, we consider baselines with different degrees of planning. We first consider binary-level planning by directly applying vanilla **DRG** in our setting: for each message, we retrieve from the generation training data the most similar utterance that has the same politeness polarity as the input message,[14] and take the strategy combination used within as the new strategy plan. We then consider a finer-grained strategy planning based on the naive **greedy** search, for which we substitute each at-risk strategy by an alternative that is the closest in strength. To make fair comparisons among different planning approaches, we apply the same set of constraints (either circumstance-induced or generation-related) we use with ILP.[15]

Evaluation. We compare the paraphrasing outputs using both automatic and human evaluations. First, we consider our *main objective*: how effective each model is at reducing the potential gap between intended and perceived politeness. We compare the predicted perceived politeness levels of paraphrases generated by each model with the intended politeness levels of the original inputs in terms of mean absolute error (**MAE** $_{gen}$), with smaller values corresponding to smaller gaps. We additionally evaluate the (pre-generation) quality of the planned strategy set (**MAE** $_{plan}$) to account for cases in which the plan is not perfectly realized.

To check the extent to which the generated paraphrases could be readily used, we assess how natural they sound to humans. We sample 100 instances from each set of the generated outputs and ask one non-author native English speaker to judge their naturalness on a scale of 1 to 5 (5 is very natural). The task is split among two annotators, and we obtain one annotation for each utterance. Each annotator was presented with an even distribution of retrieval-based, greedy-based and ILP-based generation outputs, and was not given any information on how the outputs are obtained.[16]

To validate that the original content is not drastically altered, we report BLEU scores (Papineni et al., 2002) obtained by comparing the generation outputs with the original message (BLEU-s), Additionally, we provide a rough measure of how 'ambitious' the paraphrasing plan is by counting the number of new strategies that are ADDED.

[14]We use f_{avg} to determine the binary politeness polarity.

[15]We note that even if we do not enforce these additional constraints, the baselines still under-perform the ILP solution.

[16]Table A2 in the Appendix shows the exact instructions.

	Translated communication (A)				Misaligned perceptions (B)			
	MAE $_{plan}$	**MAE** $_{gen}$	BLEU-s	#-ADDED	**MAE** $_{plan}$	**MAE** $_{gen}$	BLEU-s	#-ADDED
No intervention	0.43	0.43	64.2	0	1.01	1.01	100	0
Retrieval (DRG)	0.66	0.61	74.7	1.09	0.81	0.76	72.0	1.07
Greedy	0.35	0.35	73.5	1.20	0.48	0.47	70.3	1.82
ILP-based	**0.14**	**0.21**	67.0	2.38	**0.03**	**0.12**	68.8	2.30

Table 2: Our method is the most efficient at reducing the potential for misalignment (bolded, t-test $p < 0.001$).

	Input / Output	**Gap**
Experiment A	could you clarify what type of image is requested of centennial olympic park? thanks!	0.23
	can you clarify what type of image is requested of centennial olympic park for me ? thanks !	0.11
	where the hell did i say that? i was referring to the term 'master'.	1.30
	so where did i actually say that ? i was referring to the term 'master'.	0.70
Experiment B	thanks for accepting. how and when do we start? sorry for the late reply.	1.30
	hi, no problem. thanks for accepting. how and when do we start?	0.03
	i'd like to try out kissle, so would you please add me to [it]? thanks.	1.06
	hi ! i ' d like to try out kissle , so will you just add me to [it]?	0.01
Error case	hi, would you please reply to me at the article talk page? thanks.	0.97
	good idea . sorry , would you please reply to me at the article talk page for you ?	0.01

Table 3: Example generation outputs (we highlight the original and newly introduced markers through which the strategies are realized). For reference, we also show the (estimated) gap between the `sender`'s intention and the `receiver`'s perception after transmission. More example outputs and error cases are shown in Tables A3 and A4 in the Appendix.

Results. Table 2 shows that our **ILP-based** method is capable of significantly reducing the potential gap in politeness perceptions between the `sender` and the `receiver`, in both experiments (t-test $p < 0.001$). The comparison with the baselines underlines the virtues of supporting fine-grained planning: the effectiveness of the eventual paraphrase is largely determined by the quality of the strategy plan. This can be seen by comparing across the MAE $_{plan}$ column which shows misalignments that would result if the plans were perfectly realized. Furthermore, when planning is done too coarsely (e.g., at a binary granularity for vanilla DRG), the resultant misalignment can be even worse than not intervening at all (for translated communication).

At the same time, the paraphrases remain mostly natural, with the average annotator ratings generally fall onto 'mostly natural' category for all generation models. The exact average ratings are 4.5, 4.2, and 4.2 for the retrieval-based, greedy-based, and ILP-based generation respectively. These generation outputs also largely preserve the content of the original message, as indicated by the relatively high BLEU-s scores.[17] Considering that the ILP-based method (justifiably) implements a more ambitious plan than the baselines (compare #-ADDED), it is expected to depart more from the original input; in spite of this, the difference in naturalness is small.

5.2 Error Analysis

By inspecting the output (examples in Tables 3, A3 and A4), we identify a few issues that are preventing the model to produce ideal paraphrases, opening avenues for future improvements:

Available strategies. Between the two experimental conditions reported in Table 2, we notice that the performance (**MAE** $_{gen}$) is worse for the case of translated communication. A closer analysis reveals that this is mostly due to a particularly hard-to-replace at-risk strategy, Swearing, which is one of the few available strategies that have strong negative politeness valence. The strategy set we operationalize is by no means exhaustive. Future work

[17] As a comparison point, we note that the outputs of all methods have higher BLEU-s scores than the back-translations. We have also verified that the generated paraphrases preserve more than 90% of the non-marker tokens, further suggesting the degree of content preservation.

can consider a more comprehensive set of strategies, or even individualized collections, to allow more diverse expressions.

Capability of the generation model. From a cursory inspection, we find that the generation model has learned to incorporate the planned strategies, either by realizing simple maneuvers such as *appending* markers at sentence boundaries, to the more complex actions such as *inserting* relevant markers in reasonable positions within the messages (both exemplified in Table 3). However, the generation model does not always fully execute the strategy plan, and can make inappropriate insertions, especially in the case of the more ambitious ILP solutions. We anticipate more advanced generation models may help further improve the quality and naturalness of the paraphrases. Alternatively, dynamically integrating the limitations of the generation model as explicit planning constraints might lead to solutions that are easier to realize.

6 Discussion

In this work, we motivate and formulate the task of circumstance-sensitive intention-preserving paraphrasing and develop a methodology that shows promise in helping people more accurately communicate politeness under different communication settings. The results and limitations of our method open up several natural directions for future work.

Modeling politeness perceptions. We use a simple linear regression model to approximate how people internally interpret politeness and restrict our attention to only the set of local politeness strategies. Future work may consider more comprehensive modeling of how people form politeness perceptions or obtain more reliable causal estimates for strategy strength (Wang and Culotta, 2019).

Task formulation. We make several simplifying assumptions in our task formulation. First, we focus exclusively on a gradable stylistic aspect that is mostly decoupled from the content (Kang and Hovy, 2019), reducing the complexity required from both the perception and the generation models. Future work may consider more complex stylistic aspects and strategies that are more tied to the content, such as switching from active to passive voice. Second, we consider binary channel constraints, but in reality, the channel behavior is often less clear-cut. Future work can aim to propose more general formulations that encapsulate more properties of the circumstance.

Forms of assistance. While we have focused on offering paraphrasing options as the form of assistance, it is not the only type of assistance possible. As our generation model may not (yet) match the quality of human rewrites, there can be a potential trade-off. While an entirely automatic assistance option may put the least cognitive load on the user, it may not produce the most natural and effective rewrite, which may be possible if humans are more involved. Hence, while we work towards providing fully automated suggestions, we might also want to utilize the language ability humans possess and consider assistance approaches in the form of interpretable (partial) suggestions.

Evaluation. In our experiments, we have relied exclusively on model predictions to estimate the level of misalignment in politeness perceptions. Given the fine-grained and individualized nature of the task, using humans to ascertain the politeness of the outputs would require an extensive and relatively complex annotation setup (e.g., collecting fine-grained labels from annotators with known backgrounds for training and evaluating individualized perception models). Furthermore, to move towards more practical applications, we would also need to conduct communication-based evaluation (Newman et al., 2020) in addition to annotating individual utterances. Future work can consider adapting experiment designs from prior work (Gao et al., 2015; Hohenstein and Jung, 2018) to establish the impact of offering such intention-preserving paraphrases in real conversations, potentially by considering downstream outcomes.

Bridging the gaps in perceptions. While we focus on politeness strategies, they are not the only circumstance-sensitive linguistic signals that may be lost or altered during transmission, nor the only type that are subject to individual or cultural-specific perceptions. Other examples commonly observed in communication include, but are not limited to, formality (Rao and Tetreault, 2018) and emotional tones (Chhaya et al., 2018; Raji and de Melo, 2020). As we are provided with more opportunities to interact with people across cultural and language barriers, the risk of misunderstandings in communication also grows (Chang et al., 2020a). Thus, it is all the more important to develop tools to mitigate such risk and help foster mutual understandings.

Acknowledgments. We thank Jonathan P. Chang, Caleb Chiam, Hajin Lim, Justine Zhang, and the reviewers for their helpful comments, Kim Faughnan and Ru Zhao for the naturalness annotations. We also thank Sasha Draghici for showing us his extensive LEGO collection, which was an inspiration for the analogy used in this paper. This research was supported in part by NSF CAREER award IIS-1750615 and NSF Grant IIS-1910147.

References

Abigail Allison and Karol Hardin. 2020. Missed Opportunities to Build Rapport: A Pragmalinguistic Analysis of Interpreted Medical Conversations with Spanish-Speaking Patients. *Health Communication*, 35(4).

Dean C. Barnlund. 2017. A Transactional Model of Communication. In *Communication Theory*. Routledge.

Roy F. Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D. Vohs. 2001. Bad is stronger than good. *Review of General Psychology*, 5(4).

David K. Berlo. 1960. *The Process of Communication: An Introduction to Theory and Practice*. Holt, Rinehart and Winston.

Douglas Biber. 1988. *Variation Across Speech and Writing*. Cambridge University Press.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press.

Moira Burke and Robert Kraut. 2008. Mind Your Ps and Qs: The Impact of Politeness and Rudeness in Online Communities. In *Proceedings of CSCW*.

Jonathan P. Chang, Justin Cheng, and Cristian Danescu-Niculescu-Mizil. 2020a. Don't Let Me Be Misunderstood:Comparing Intentions and Perceptions in Online Discussions. In *Proceedings of WWW*.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020b. ConvoKit: A Toolkit for the Analysis of Conversations. In *Proceedings of SIGDIAL*.

Niyati Chhaya, Kushal Chawla, Tanya Goyal, Projjal Chanda, and Jaya Singh. 2018. Frustrated, Polite, or Formal: Quantifying Feelings and Tone in Email. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*.

Elizabeth Clark, Anne S. Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In *Proceedings of IUI*.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A Computational Approach to Politeness with Application to Social Factors. In *Proceedings of ACL*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In *Prceedings of ICLR*.

Jessica Ficler and Yoav Goldberg. 2017. Controlling Linguistic Style Aspects in Neural Language Generation. In *Proceedings of the Workshop on Stylistic Variation*.

Joseph P. Forgas. 1998. Asking Nicely? The Effects of Mood on Responding to More or Less Polite Requests. *Personality and Social Psychology Bulletin*, 24(2).

Yao Fu, Yansong Feng, and John P Cunningham. 2019. Paraphrase Generation with Latent Bag of Words. In *Proceeding of NeurIPS*.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style Transfer in Text: Exploration and Evaluation. In *Proceedings of AAAI*.

Ge Gao, Bin Xu, David C. Hau, Zheng Yao, Dan Cosley, and Susan R. Fussell. 2015. Two is Better Than One: Improving Multilingual Collaboration by Giving Two Machine Translation Outputs. In *Proceeding of CSCW*.

Jeffrey T. Hancock, Mor Naaman, and Karen Levy. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication*, 25(1).

Susan C. Herring. 1994. Politeness in Computer Culture: Why Women Thank and Men Flame. *Cultural Performances: Proceedings of the Third Berkeley Women and Language Conference*.

Susan C. Herring. 1996. *Computer-Mediated Communication: Linguistic, Social, and Cross-Cultural Perspectives*. John Benjamins Publishing.

Jess Hohenstein and Malte Jung. 2018. AI-Supported Messaging: An Investigation of Human-Human Text Conversation with AI Support. In *Proceedings of CHI. Extended Abstract.*

Yuheng Hu, Ali Tafti, and David Gal. 2019. Read This, Please? The Role of Politeness in Customer Service Engagement on Social Media. In *Proceedings of HICSS*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward Controlled Generation of Text. In *Proceedings of ICML*.

Yiqing Hua, Cristian Danescu-Niculescu-Mizil, Dario Taraborelli, Nithum Thain, Jeffery Sorensen, and Lucas Dixon. 2018. WikiConv: A Corpus of the Complete Conversational History of a Large Online Collaborative Community. In *Proceedings of EMNLP*.

Dongyeop Kang and Eduard Hovy. 2019. xSlue: A Benchmark and Analysis Platform for Cross-Style Language Understanding and Evaluation. *arXiv:1911.03663v1 [cs.CL]*.

Gabriele Kasper. 1990. Cross-Cultural Perspectives on Linguistic Politeness. *Journal of Pragmatics*, 14(2).

Nitish S. Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A Conditional Transformer Language Model for Controllable Generation. *arXiv:1909.05858v2 [cs.CL]*.

Robin T. Lakoff. 1973. *The Logic of Politeness: Minding Your P's and Q's*. Chicago Linguistic Society.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, Retrieve, Generate: A Simple Approach to Sentiment and Style Transfer. In *Proceedings of ACL*.

Xiangyang Li, Guo Pu, Keyu Ming, Pu Li, Jie Wang, and Yuxuan Wang. 2020. Review of Text Style Transfer Based on Deep Learning. *arXiv:2005.02914v1 [cs.CL]*.

Yossi Maaravi, Orly Idan, and Guy Hochman. 2019. And Sympathy Is What We Need My Friend—Polite Requests Improve Negotiation Results. *PLoS One*, 14(3).

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W. Black, and Shrimai Prabhumoye. 2020. Politeness Transfer: A Tag and Generate Approach. In *Prceedings of ACL*.

Hiroko Matsuura. 1998. Japanese EFL Learners' Perception of Politeness in Low Imposition Requests. *JALT Journal*, 20(1).

Marie Meteer and Varda Shaked. 1988. Strategies for Effective Paraphrasing. In *Proceedings of COLING*.

Hannah J. Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. 2016. "Blissfully Happy" or "Ready to Fight": Varying Interpretations of Emoji. In *Proceedings of ICWSM*.

Stuart Mitchell, Michael O'Sullivan, and Iain Dunning. 2011. PuLP: A Linear Programming Toolkit for Python.

Seyed I. Mousavi and Reza G. Samar. 2013. Contrastive Rhetoric: Investigating Politeness and Intimacy in Business Email Communications in Four Asian Countries. *The International Journal of Humanities*, 19(1).

James Murphy. 2014. (Im)politeness During Prime Minister's Questions in the U.K. Parliament. *Pragmatics and Society*, 5(1).

Courtney Napoles, Maria Nădejde, and Joel Tetreault. 2019. Enabling Robust Grammatical Error Correction in New Domains: Data Sets, Metrics, and Analyses. *TACL*, 7.

Benjamin Newman, Reuben Cohn-Gordon, and Christopher Potts. 2020. Communication-based Evaluation for Natural Language Generation. In *Proceedings of SCiL*.

Tong Niu and Mohit Bansal. 2018. Polite Dialogue Generation Without Parallel Data. *TACL*, 6.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. 2018. Style Transfer Through Back-Translation. In *Proceedings of ACL*.

Chris Quirk, Chris Brockett, and William Dolan. 2004. Monolingual Machine Translation for Paraphrase Generation. In *Proceeding of EMNLP*.

Shahab Raji and Gerard de Melo. 2020. What Sparks Joy: The AffectVec Emotion Database. In *Proceedings of WWW*.

Sudha Rao and Joel Tetreault. 2018. Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer. In *Proceedings of NAACL*.

Philip Riley. 1984. Understanding Misunderstandings: Cross-Cultural Pragmatic Failure in the Language Classroom. *European Journal of Teacher Education*, 7(2).

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling Politeness in Neural Machine Translation via Side Constraints. In *Proceeding of NAACL*.

Claude E. Shannon and Warren Weaver. 1963. *The Mathematical Theory of Communication*. University of Illinois Press.

Matthew Shardlow. 2014. A Survey of Automated Text Simplification. In *Proceeding of IJACSA*.

Brandon Stevens. 2019. Grammar? Check: A Look at Grammar-Checking Software in the Context of Tutoring. In *Prceedings of ECWCA*.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. Transforming Delete, Retrieve, Generate Approach for Controlled Text Style Transfer. In *Proceedings of EMNLP*.

Jenny Thomas. 1983. Cross-Cultural Pragmatic Failure. *Applied Linguistics*, 4(2).

Sergiy Tyupa. 2011. A Theoretical Framework for Back-Translation as a Quality Assessment Tool. *New Voices in Translation Studies*, 7(1).

John Velentzas and Georgia Broni. 2014. Communication Cycle: Definition, Process, Models and Examples. *Recent Advances in Financial Planning and Product Development*.

Rob Voigt, Nicholas P. Camp, Vinodkumar Prabhakaran, William L. Hamilton, Rebecca C. Hetey, Camilla M. Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L. Eberhardt. 2017. Language from Police Body Camera Footage Shows Racial Disparities in Officer Respect. *PNAS*, 114(25).

Zhao Wang and Aron Culotta. 2019. When do Words Matter? Understanding the Impact of Lexical Choice on Audience Perception using Individual Treatment Effect Estimation. In *Proceedings of AAAI*.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for Style. In *Proceedings of COLING*.

Michael Yeomans, Alejandro Kantor, and Dustin Tingley. 2019. The Politeness Package: Detecting Politeness in Natural Language. *The R Journal*, 10(2).

Yiheng Zhou, He He, Alan W. Black, and Yulia Tsvetkov. 2019. A Dynamic Strategy Coach for Effective Negotiation. In *Proceedings of SIGDIAL*.

Appendices

A Politeness Strategies

We show the complete list of politeness strategies we use in Table A1, together with the coefficients for the average model f_{avg} used in Experiment A is shown in Table A1.

Recognizing that individual markers may not always fully encompass the politeness-bearing portion of the text, we consider two modes of deletion depending on strategy (Table A1): *token* mode deletes only the identifier marker, whereas in *segment* mode the whole sentence segment (as defined by within-sentence punctuations) will be removed:

Token mode	Can you ~~**please**~~ explain?
Segment mode	~~**Thanks** for your help,~~ I will try again.

B Prolific Annotators

For experiment B, we sample the top five most prolific annotations from the Wikipedia section of the Stanford Politeness Corpus, with the most prolific one having annotated 2,063 instances, and the least prolific among the five having 715 annotations.

When training individual perception models, we note that some less frequently used strategies tend to be under annotated at the individual level, and may thus create artificially high difference in coefficients. We thus use the coefficient from the average model for any strategy that is annotated for less than 15 times by the individual annotator.

C Additional Details on ILP

We consider a few linguistic constraints to help exclude some counter-intuitive strategy combinations. It should be noted that, with increased quality of a generation model, or by dynamically integrating the limitation of the generation model into the planning step, the process of inserting such additional constraints may be automated:

Negativity constraint. While our simple linear model estimates the level of politeness by the aggregated effects of all strategies used regardless of their polarity, humans are known to have a negativity bias (Baumeister et al., 2001): while the presence of polite markers in an otherwise impolite utterance may soften the tone, the use of a negative marker in an otherwise polite utterance may be overshadowing. As a result, when an input is judged to be positive in politeness, we consider the additional constraint to exclude use of negative strategies, i.e., $x_s = 0, \forall s \in \{s : b_s < 0\}$.

Subjunctive and Indicative constraint. Admittedly, among the set of markers we consider, some are more decoupled from contents than others—while removing *just* is almost guaranteed to keep the original meaning of the sentence intact, for an utterance that starts with either Subjunctive or Indicative, e.g., *could you clarify?*, simply removing *could you* would have already made its meaning ambiguous.[18] To account for this, we add the constraint that the use of Subjunctive and Indicative should be substituted within themselves, i.e., $x_{\text{Subjunctive}} + x_{\text{Indicative}} = \mathbb{1}_{\mathcal{S}_{in}(\text{Subjunctive})} + \mathbb{1}_{\mathcal{S}_{in}(\text{Indicative})}$.[19]

D Details on Human Evaluations

To evaluate on the naturalness of the generated text, we ask two non-author native speaker for naturalness ratings on a scale of 1 (very unnatural) to 5 (very natural). The exact instruction is shown in Table A2.

[18] For instance, *can I clarify?* and *can you clarify?* would both be linguistically plausible requests containing *clarify*, yet they differ significantly in meaning.

[19] We acknowledge that under certain circumstances, this constraint may be impossible to fulfill.

Strategy	Coeff.	Example markers	Delete mode	Example usage
Actually	-0.358	really, actually	token	it **actually** needs to be ...
Adverb.Just	-0.004	just	token	i **just** noticed that ...
Affirmation	0.171	ok, good [work]	segment	**excellent point**, i have added it ...
Apology	0.429	sorry, [i] apologize	segment	**sorry** to be off-topic but ...
By.The.Way	0.331	by the way, btw	token	okay - **btw**, do you want me ...?
Conj.Start	-0.245	so, and, but	token	**so** where is the article ?
Filler	-0.245	hmm, um	token	**uh**, hey, can you...?
For.Me	0.128	for me	token	is it alright **for me** to archive it now?
For.You	0.197	for you	token	i can fetch one **for you** in a moment! ...
Gratitude	0.989	thanks, [i] appreciate	segment	**thanks** for the info , ...
Greeting	0.491	hi, hello	token	**hey** simon , help is needed if possible ...
Hedges	0.131	possibly, maybe	token	**maybe** some kind of citation is needed ...
Indicative	0.221	can you, will you	token	**can you** create one for me?
Please	0.230	please	token	can you **please** check it?
Please.Start	-0.209	please	token	**please** stop . if you continue ...
Reassurance	0.668	no worries	segment	**no problem**, happy editing. ...
Subjunctive	0.454	could you, would you	token	..., **could you** check?
Swearing	-1.30	the hell, fucking	token	what **the heck** are you talking about?

Table A1: Local politeness strategies being considered. For each strategy, we show its corresponding coefficients in the linear regression model, example markers, together with example usages.

Ignoring punctuations, typos, and missing context, on a scale of 1-5, how natural does the text sound?
5. **Very natural**: It's possible to imagine a native speaker sending the message online.
4. **Mostly natural**: While there are some minor errors, simple edits can make it become 'very natural'.
3. **Somewhere in between**: While the text is comprehensible, it takes more involved edits to make it sound natural.
2. **Mostly unnatural**: There are significant grammatical issues that make the text almost not comprehensible.
1. **Very unnatural**: Entirely broken English.

Table A2: Instruction for naturalness annotations.

E Additional Generation Examples

We show additional generation outputs in Table A3, and a categorization of failure cases in Table A4.

Strategy plan	Input (upper) / Output (lower)	Score
Please, Subjunctive, Gratitude	could you then please make some contributions in some of your many areas of expertise? thanks.	
Greeting, Subjunctive, Adverb.Just, For.Me	hi , could you then just make some contributions for me in some of your many areas of expertise ?	5
Please	can someone please explain why there's a coi tag on this article? it's not evident from the talk page.	
For.Me, Hedges	can someone explain why there ' s a coi tag on this article for me ? it ' s not apparent from the talk page .	5
Conj.Start, Filler	uh...ok...whatever...did you get that user name yet? or do you prefer hiding behind your ip?	
Actually, By.The.Way, Conj.Start, Please.Start	ok . . . whatever . . . did you actually get that user name yet ? or do you prefer hiding behind your ip ?	5
Please, Subjunctive	could you please stop your whining, and think about solutions instead? tx.	
By.The.Way, Hedges, Indicative	btw , can you maybe stop your whining , and think about solutions instead ? tx .	5

Table A3: Additional examples from the generation outputs, together with strategy information (*original strategy combination* for inputs in italics, realized strategies underlined for outputs) and naturalness scores. We also highlight the original and newly introduced markers through which the strategies are realized. Refer to Table A4 for common types of failure cases.

Error type	Input (upper) / Output (lower)	Score
`Grammatical mistake`	the bot seems to be down again. could you give it a nudge? the bot seems to be down again . **maybe** can you give it a nudge for me ?	3
	i see you blocked could you provide your rationale? thanks - () i see you blocked **provide** your rationale ? (**please**)	2
`Strategy misfit`	hello, this image has no license info, could you please add it? thank you. hello , this image has no license info , **sorry** . could you add it **for you** ? thank you .	3
	can you please review this or block or have it reviewed at ani? thank you **no worries . sorry** , can you review this or block or have it reviewed for me at ani ?	3

Table A4: Examples demonstrating two representative error types with naturalness scores. `Grammatical mistake` represents cases when the markers are in inappropriate positions or introduce errors to the sentence structure. `Strategy misfit` represents cases when the use of suggested strategies (regardless of choice of markers to realize them) do not seem appropriate. Problematic portions of the outputs are in bold.