CaSiNo: A Corpus of Campsite Negotiation Dialogues for Automatic Negotiation Systems

Kushal Chawla[1]   **Jaysa Ramirez**[2*]   **Rene Clever**[3*]   **Gale Lucas**[1]
Jonathan May[4]   **Jonathan Gratch**[1]

[1&4]University of Southern California, Los Angeles, USA
[2]Rollins College, Winter Park, USA   [3]CUNY Lehman College, Bronx, USA
[1]{chawla,lucas,gratch}@ict.usc.edu
[2]jramirez@rollins.edu   [3]rene.clever@lc.cuny.edu
[4]jonmay@isi.edu

Abstract

Automated systems that negotiate with humans have broad applications in pedagogy and conversational AI. To advance the development of practical negotiation systems, we present CaSiNo: a novel corpus of over a thousand negotiation dialogues in English. Participants take the role of campsite neighbors and negotiate for food, water, and firewood packages for their upcoming trip. Our design results in diverse and linguistically rich negotiations while maintaining a tractable, closed-domain environment. Inspired by the literature in human-human negotiations, we annotate persuasion strategies and perform correlation analysis to understand how the dialogue behaviors are associated with the negotiation performance. We further propose and evaluate a multi-task framework to recognize these strategies in a given utterance. We find that multi-task learning substantially improves the performance for all strategy labels, especially for the ones that are the most skewed. We release the dataset, annotations, and the code to propel future work in human-machine negotiations: https://github.com/kushalchawla/CaSiNo.

1 Introduction

Negotiations are highly prevalent in our interactions, from deciding who performs the household chores to high-stake business deals to maintaining international peace. Automatic negotiation systems are helpful in providing cost-effective social skills training (Johnson et al., 2019) and for advanced capabilities of AI assistants such as Google Duplex (Leviathan and Matias, 2018).

A negotiation requires understanding the partner's motives along with effective reasoning and communication, which is challenging for an automated system. Prior work in human-machine negotiations primarily uses strict communication protocols such as a pre-defined menu of options (Mell

and Gratch, 2016). Systems involving free-form dialogue are limited due to a lack of interdisciplinary efforts in NLP and Computational Social Science in this direction. Initial efforts in building dialogue systems for negotiations looked at game environments (Asher et al., 2016; Lewis et al., 2017). DealOrNoDeal (Lewis et al., 2017) involves two negotiators who split given quantities of three arbitrary items: books, balls, and hats. This provides a concrete structure to the negotiation, keeps the design tractable, and ensures a reliable evaluation based on final points scored. Many practical solutions in negotiations follow similar *closed-domain* designs (Mell and Gratch, 2016). However, most of the dialogues in these game settings reduce to merely an exchange of offers from both sides. For instance, '*i need the book and the balls you can have the hat*' or '*i want the ball and 2 books*' in DealOrNoDeal. One reason for this lack of richness in language use is that the items are *arbitrarily defined*, that is, there is *no semantic context* around the items that the participants are negotiating for. Hence, this setup fails to capture many realistic aspects of negotiations such as small talk, preference elicitation, emotion expression, and convincing strategies based on individual preferences and requirements. Emulating real-world negotiations is desirable for developing practical systems for social skills training and robust AI assistants that are useful in realistic scenarios.

On the other extreme, the CB dataset (He et al., 2018) involves buyer-seller negotiations to finalize the price of a given product. Targeting the collection of more *open-ended* dialogues, the participants are also encouraged to discuss side offers, such as free delivery or also selling other accessories at the same price. Although this promotes diversity and rich natural conversations, unfortunately, such open-ended domains make the evaluation of negotiation performance non-trivial, which also inhibits the practical applicability of the systems developed

---

*Work done when authors were interns at USC ICT

3167

on such datasets. For instance, in skills training, it is desirable to judge the performance and provide critical feedback (Monahan et al., 2018).

To address these shortcomings, we design a novel negotiation task. Our design is based on a tractable *closed-domain* abstraction from the negotiation literature but is infused with a real-world camping scenario, resulting in rich dialogues for natural language research (Section 2). The task involves two participants who take the role of campsite neighbors and negotiate for additional *Food*, *Water*, and *Firewood*, based on individual preferences and requirements.

Based on this design, we collect **CaSiNo**: a corpus of 1030 **Ca**mp **Si**te **Neg**otiation dialogues in English. The dialogues contain various aspects of a realistic negotiation, such as rapport building, discussing preferences, exchanging offers, emotion expression, and persuasion with personal and logical arguments. We also collect the participants' satisfaction from the outcome and how much they like their opponents, both being important metrics in negotiations (Mell et al., 2019). We annotate 9 persuasion strategies that span cooperative to selfish dialog behaviors (Section 3). We perform an extensive correlational analysis to investigate the relationship among the final outcomes and explore how they relate to the use of negotiation strategies (Section 4). Further, we propose a multi-task framework with task-specific self-attention mechanisms to recognize these strategies in a given utterance (Section 5). Our insights form the foundation for the development of practical negotiation systems that engage in free-form natural conversations. We release the dataset along with the annotations to enable future work in this direction.

2   The CaSiNo Dataset

Our data was crowd-sourced on Amazon Mechanical Turk. We describe our design by following the journey of a specific participant in our study.

Pre-Survey: We start by collecting demographics and psychological personality traits of the participants which relate to their negotiation behaviors. For demographics, we gather age, gender, ethnicity, and the highest level of education. We consider two measures of individual personality differences: Social Value Orientation or SVO (Van Lange et al., 1997) and Big-5 personality (Goldberg, 1990) that have been heavily studied in the context of negotiations (Bogaert et al., 2008; Curtis et al., 2015).

SVO classifies the participants as *Prosocial*, who tend to approach negotiations cooperatively, or *Proself*, who tend to be more individualistic. Big-5 personality test assesses the participants on five dimensions: Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experiences. Our participants exhibit diverse demography and psychological personality. We provide aggregate statistics in Appendix A.

Negotiation Training: Research shows that the average human is bad at negotiating (Wunderle, 2007; Babcock and Laschever, 2009), which can adversely impact the quality of the collected dialogues and consequently, the system trained on them. One way to mitigate this is by using reinforcement learning to optimize on a reward that measures the negotiation performance. RL training has proved to be challenging and often leads to degeneracy (Lewis et al., 2017). Further, this ignores prior work in human-human negotiations that provides guidelines for achieving favorable outcomes in realistic negotiations (Lewicki et al., 2016).

To incorporate these best practices in a principled way, we design a training module. Each participant is asked to watch a video tutorial before their negotiation. The tutorial takes an example of a negotiation between two art collectors to encourage them to follow some of the best practices in negotiations (Lewicki et al., 2016), including 1) Starting with high offers, 2) Discussing preferences, 3) Appropriate emotion expression, and 4) Discussing individual requirements to make convincing arguments. This results in a rich and diverse set of dialogues, as we explore further in later sections. We release the complete video tutorial publicly, with the hope that it promotes reproducibility and helps researchers to design similar data collection experiments in the future: https://youtu.be/7WLy8qjjMTY.

Preparation Phase: Several requirements guide our design choices: 1) *Semantically Meaningful*: The context must be meaningful and relatable for MTurk participants and for anyone who negotiates with the system trained on this dataset. This allows the participants to indulge in personal and contextual conversations, making the resulting system more useful for downstream applications. 2) *Symmetric task*: The task should be symmetric for both the participants so that a dialogue system may leverage both sides of the conversations during modelling, and 3) *Symmetric items*: The items

Preferences & Arguments	
P1	**P2**
High: Water: We like to go on runs and it increases the need of this.	*High*: Food: Food really increases everyones morale.
Medium: Food: Food overall is a good mood booster.	*Medium*: Firewood: We like to have a large fire.
Low: Firewood: We do not care for fire and it is not necessary to us.	*Low*: Water: We don't drink water that often.

Conversation	Annotation
P1: How are you today? Did you have any preferences on the supplies we will be trading?	Small-Talk, Coordination, Elicit-Pref
P2: I am good. How about yourself? I think I would like some firewood to start off with. We like to have bigger fires. What about you?	Small-Talk, Self-Need, Other-Need, Elicit-Pref
P1: I am good as well. That is good to hear that you like to have bigger fires as we do not care much for that. We would much rather have some extra water.	Small-Talk, Empathy, No-Need
P2: Water is a little important to us too though , if possible maybe we can split that or maybe we can get some more food in replacement.	Coordination
P1: That may be possible.... 🙂 What did you have in mind for the food replacement?	Non-strategic
P2: You can have all the water if we can have all the food?	Non-strategic
P1: I dont think I am okay with that ☹ . Food is essential to our groups morale when camping. We would like 1 additional food preferably.	Self-Need, Other-Need
P2: Well you guys did say you did not care much about large fires. What if you gave all the firewood in replace for the water and you can still keep 1 food?	UV-Part, Coordination
P1: So I would get 3 water and 1 food and youd get 3 firewood and 2 food?	Non-strategic
P2: Yea that seems like an alright trade to me 🙂	Non-strategic
P1: Hmm... alright then 🙂	Non-strategic
P2: Submit-Deal	
P1: Accept-Deal	

Table 1: Sample dialogue from the CaSiNo dataset. **P1** and **P2** represent two participants in our study.

which the participants are negotiating for should be symmetric in the sense that an individual can resonate with any preference order assigned to them. Hence, every category of items can be more desirable over others depending on a real-world context.

Our scenario is an instance of a common and useful abstraction for studying negotiations in scientific literature known as the multi-issue bargaining task (Fershtman, 1990). The task involves campsite neighbors who negotiate for additional *Food*, *Water*, and *Firewood* packages, each with a total quantity of three. Instead of choosing an arbitrary set of items, each item represents quite relatable, basic requirements that one might plausibly have for an actual camping trip. The items were only broadly defined to encourage diversity. One challenge when dealing with a realistic context like camping is the inherent bias that one might have towards one item over others, which violates our symmetry constraint. To mitigate this, we emphasize that the camping authorities have already provided the basic essentials and the participants will be negotiating for extras, based on their individual plans for camping. We present the negotiation

scenario, as seen by participants, in Appendix B.

The three item types are assigned a random priority order for every participant using a permutation of {*High*, *Medium*, *Low*}. As in realistic negotiations, the participants are asked to *prepare* for their negotiation by coming up with justifications for the given preferences before the negotiation begins (precise question format in Appendix G), for instance, needing more water supplies for a hike or firewood for a bonfire with friends. We find that the participants are able to come up with a variety of arguments from their own camping experiences, such as *Personal Care*, *Recreational*, *Group Needs* or *Emergency* requirements. We illustrate some of these arguments in Appendix B. The participants were encouraged to use their justifications as they feel fit, to negotiate for a more favorable deal.

Negotiation Dialogue: Finally, two participants are randomly paired to engage in an alternating dialogue for a minimum total of 10 utterances. We also provide the option to use emoticons for four basic emotions, namely, happy, sad, anger, and surprise. After coming to an agreement, the participants submit the deal formally using the provided

options. They can also walk away from the negotiation if they are unable to come to an agreement. The primary evaluation metric to assess the negotiation performance is the number of points scored by a negotiator. Every *High*, *Medium*, and *Low* priority item is worth 5, 4, and 3 points respectively, such that a participant can earn a maximum of 36 points if she is able to get *all* the available items.

Post-Survey: We collect two other evaluation metrics relevant to negotiations: 1) 5-point scale for satisfaction (How satisfied are you with the negotiation outcome?) and 2) 5-point scale for opponent likeness (How much do you like your opponent?). Back-to-back negotiation (Aydoğan et al., 2020) is an interesting case where the relationship with the partner is crucial. In such a case, a poor relationship in earlier negotiations can adversely impact the performance in later rounds. Further, for some cases in CaSiNo, we observed that the participants were satisfied with their performance, despite performing poorly because they thought that the arguments of their partners for claiming the items were justified. One might argue that this is still a successful negotiation. Hence, we believe that all the metrics defined in the paper are important in the context of real-world negotiations and propose that they should be looked at collectively. We will further analyze these outcome variables in Section 4 where we study the correlations between the participants' negotiation behaviors and these metrics of negotiation performance.

Data Collection: We collected the dataset over a month using the ParlAI framework (Miller et al., 2017). Screenshots from the interface are provided in Appendix G. The participant pool was restricted to the United States, with a minimum 500 assignments approved and at least 95% approval rate. We post-process the data to address poor quality dialogues and inappropriate language use. We describe these post-processing steps in Appendix C.

Finally, we end up with 1030 negotiation dialogues between 846 unique participants. On average, a dialogue consists of 11.6 utterances with 22 tokens per utterance. We present a sample dialogue with the associated participant profile in Table 1. The participants are rewarded a base amount of $2 for their time (around 20 minutes). Further, they were incentivized with a performance-based bonus of 8.33 cents for every point that they are able to negotiate for. If a participant walks away, both parties get the amount corresponding to one high item

or the equivalent of 5 points. The bonus is paid out immediately after the task to encourage participation. We discuss ethical considerations around our data collection procedure in Section 8. Overall, the participants had highly positive feedback for our task and could relate well with the camping scenario, engaging in enjoyable, interesting, and rich personal conversations. We discuss their feedback with examples in Appendix D.

3 Strategy Annotations

Label	Example	Count	α
	Prosocial Generic		
Small-Talk	Hello, how are you today?	1054	0.81
Empathy	Oh I wouldn't want for you to freeze	254	0.42
Coordination	Let's try to make a deal that benefits us both!	579	0.42
	Prosocial About Preferences		
No-Need	We have plenty of water to spare.	196	0.77
Elicit-Pref	What supplies do you prefer to take the most of?	377	0.77
	Proself Generic		
Undervalue-Partner	Do you have help carrying all that extra firewood? Could be heavy?	131	0.72
Vouch-Fairness	That would leave me with no water.	439	0.62
	Proself About Preferences		
Self-Need	I can't take cold and would badly need to have more firewood.	964	0.75
Other-Need	we got kids on this trip, they need food too.	409	0.89
Non-strategic	Hello, I need supplies for the trip!	1455	-

Table 2: Utterance-level strategy annotations. α refers to Krippendorff's alpha among 3 annotators on a subset of 10 dialogues ($\sim$ 120 utterances). An utterance can have multiple labels.

After collecting the dataset, we developed an annotation schema to analyze the negotiation strategies used by the participants, and to facilitate future work. We follow the conceptual content analysis procedure (Krippendorff, 2004) to design the scheme. Being a natural conversational dataset, we find several instances where a strategy spans multiple sentences in an utterance, as well as instances where the same sentence contains several strategies. Hence, we define an utterance as the *level of analysis*. Each utterance is annotated with one or more labels. If no strategy is evident, the utterance is labelled as **Non-strategic**. Although we label entire utterances, self-attention shows some promise as an automatic way to identify which part of an utterance corresponds to a given strategy, if desirable for a downstream application (Section 5).

Human negotiation behaviors can be broadly cat-

egorized as Prosocial, which promote the interests of others or the common good, and Proself, which tend to promote self-interest in the negotiations (Yamagishi et al., 2017; Van Lange et al., 2007). Another important criterion is discussing preferences. Prior work suggests that humans negotiate with a fixed-pie bias, assuming that the partner's preferences align, and hence achieving sub-optimal solutions (Kelley, 1996). Based on these distinctions and manual inspection, we define 9 strategies used in the CaSiNo dataset. The usage of these negotiation strategies correlates with both the objective and subjective metrics of negotiation performance.

3.1 Prosocial

Prosocial strategies address the concerns of both self and the negotiation partner. We define three strategies that exhibit *generic Prosocial behavior*.

Small-Talk: Participants engage in small talk while discussing topics apart from the negotiation, in an attempt to build a rapport with the partner. For example, discussing how the partner is doing during the pandemic or sharing excitement for the camping trip. Rapport has been well studied to positively impact negotiation outcomes (Nadler, 2003). Small talk usually appears either at the beginning or at the end of the negotiation.

Empathy: An utterance depicts **Empathy** when there is evidence of positive acknowledgments or empathetic behavior towards a personal context of the partner, for instance, towards a medical emergency. Empathy promotes Prosocial behaviors in interpersonal interactions (Klimecki, 2019).

Coordination is used when a participant promotes coordination among the two partners. This can be, for instance, through an explicit offer of a trade or mutual concession, or via an implicit remark suggesting to work together towards a deal. Further, we define two strategies that relate to *Prosocial behavior about individual preferences*:

No-Need is when a participant points out that they do not need an item based on personal context such as suggesting that they have ample water to spare. **No-Need** can directly benefit the opponent since it implies that the item is up for grabs.

Elicit-Pref is an attempt to discover the preference order of the opponent. CaSiNo covers a range of scenarios based on how aligned the preferences of the two parties are. Generally, we find that discussing preferences upfront leads to smoother negotiations without much back and forth.

3.2 Proself

Proself behavior attempts to serve personal performance in a negotiation. We define two strategies exhibiting *generic Proself behavior*.

Undervalue-Partner or **UV-Part**, refers to the scenario where a participant undermines the requirements of their opponent, for instance, suggesting that the partner would not need more firewood since they already have the basic supplies or a suggestion that there might be a store near the campsite where the partner can get the supplies instead.

Vouch-Fairness is a callout to fairness for personal benefit, either when acknowledging a fair deal or when the opponent offers a deal that benefits them. For instance, through an explicit callout 'this deal is not fair', or implicitly saying 'this does not leave me with anything'.

Finally, we consider two *Proself strategies that relate to individual preferences*:

Self-Need refers to arguments for creating a personal need for an item in the negotiation. For instance, a participant pointing out that they sweat a lot to show preference towards water packages.

Other-Need is similar to **Self-Need** but is used when the participants discuss a need for someone else rather than themselves. For instance, describing the need for firewood to keep the kids warm. Negotiating on behalf of others is densely studied as a competitive strategy, where negotiators engage in contentious, demanding, and inflexible bargaining behaviors (Adams, 1976; Clopton, 1984).

Collecting annotations: Three expert annotators[1] independently annotated 396 dialogues containing 4615 utterances. The annotation guidelines were iterated over a subset of 5 dialogues, while the reliability scores were computed on a different subset of 10 dialogues. We use the nominal form of Krippendorff's alpha (Krippendorff, 2018) to measure the inter-annotator agreement. We provide the annotation statistics in Table 2. Although we release all the annotations, we skip **Coordination** and **Empathy** for our analysis in this work, due to higher subjectivity resulting in relatively lower reliability scores. For the rest of the paper, we will refer to this annotated subset of CaSiNo as CaSiNo-Ann.

4 Correlational Analysis

We next perform correlational analysis on CaSiNo-Ann to understand how the points scored by a participant relate to their satisfaction from the outcome

---

[1]Researchers involved in the project.

and their opponent perception. We further shed light on what kind of strategies are more likely to lead to better outcomes. Such insights motivate our experiments on strategy prediction and would direct future efforts in building negotiation systems. We present complete results in Appendix E and discuss the significant observations below.

Relationship among outcome variables: We consider the points scored, satisfaction from the outcome, and opponent likeness. We find that the points scored by a participant are positively correlated with their own satisfaction (r=0.376, $p < 0.01$) and with their perception of the opponent (r=0.276, $p < 0.01$). Similar trends are visible with the corresponding variables of the negotiation partner as well, suggesting that the participants secured more points while still maintaining a positive perception in the eyes of their opponents.

Discovering the integrative potential: Integrative potential in a negotiation is based on how aligned the partner preferences are. Complete alignment leads to a *distributive* (or zero-sum) negotiation, having a low integrative potential where the benefit of one results in a high loss for the other. A negotiation is *integrative* if the preferences do not align, allowing for solutions that maximize mutual points. We assign each dialogue either 1, 2, or 3, depending on whether the integrative potential is *low*, *medium*, or *high*. The maximum joint points possible in these cases are 36, 39, and 42 respectively. We find that the participants are able to discover this integrativeness, thereby achieving significantly more joint points as the potential increases ($r = 0.425$, $p < 0.001$).

Use of negotiation strategies: Overall, we find that greater use of Prosocial strategies shows a general pattern to predict higher ratings for both subjective measures of satisfaction and likeness, for self as well as the partner. Engaging in small talk shows significant positive correlations ($ps < 0.01$), confirming our hypothesis from prior work that it relates to healthier relationships among the negotiators. Similar effects are visible for **No-Need** ($ps < 0.05$), where the participant decides to let go one of their low-priority items. Since this directly benefits the opponent, it is likely to improve the participant's perception. On the other hand, Proself strategies show a general pattern to predict lower satisfaction and likeness ratings for both self and the partner. We observe significant negative correlation for both **Other-Need** and **Vouch-Fair**

($ps < 0.01$). Further, we find that these competitive strategies are also associated with lower points scored by the participant and the opponent, and hence, the joint points ($ps < 0.01$). These correlations are not influenced by the integrative potential in the scenario, as when the integrated potential is controlled for, the effects generally remain unchanged and demonstrate the same patterns.

We further observe that the dialogue behavior of a negotiator significantly relates to the behavior of their opponent, where both tend to use similar negotiation strategies ($ps < 0.01$). Our findings show that Prosocial strategies are more likely to be associated with Prosocial behavior in the opponents and achieve more favorable outcomes in our negotiation scenario as compared to Proself. These results suggest that an automated negotiator can benefit by employing different strategies based on Prosocial or Proself behaviors of the opponent, for instance, by matching Prosocial behaviors but not Proself. The first step in this direction is to recognize them in a given utterance, which is our focus in the next section.

5 Strategy Prediction

For building an automated dialogue system that incorporates the negotiation strategies discussed above, an important first step is to build computational models that recognize their usage in the observed utterances. Hence, we explore the task of strategy prediction, given an utterance and its previous dialogue context.

5.1 Methodology

Pre-trained models have proved to be useful on a number of supervised tasks with limited in-domain datasets. Inspired by this success, we use BERT-base (Devlin et al., 2019) as the core encoding module. A natural way to use pre-trained models for our task is to fine-tune the model for every label independently in a binary classification setup, where the positive class represents the presence of a strategy, and the negative represents its absence. However, most of the utterances in the CaSiNo-Ann dataset are **Non-strategic**, resulting in a high imbalance where most of the data points belong to the negative class. As we later show, directly fine-tuning the BERT model fails to recognize the strategies for which the data is most skewed.

We instead propose a multi-task learning framework to allow parameter sharing between the dif-
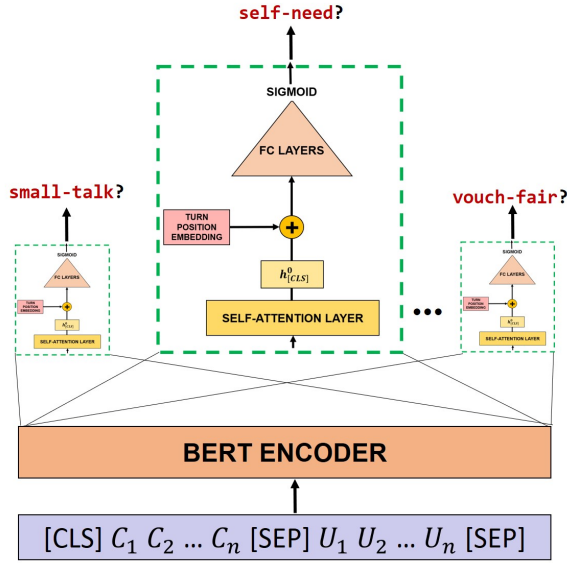
Figure 1: Architecture for multi-task strategy prediction. **+** represents element-wise summation.

ferent prediction tasks. Our architecture involves a common BERT-base encoder shared with all the tasks but uses task-specific self-attention to allow the model to focus on the most relevant parts of the input for each task separately. Consequently, this also enables interpretability by allowing us to visualize which parts of an utterance are attended for any given strategy. Our input consists of a finite size context window, which loses the turn index for a specific utterance. Hence, we also capture the turn position for each utterance using sinusoidal positional embeddings (Vaswani et al., 2017). We present the complete architecture in Figure 1.

In-Domain Pre-Training (IDPT): CaSiNo-Ann is nearly 40% of the entire CaSiNo dataset. To incorporate the unannotated dialogues, we employ In-Domain Pre-training of the BERT encoder (Sun et al., 2019). For this purpose, we consider each unannotated dialogue as a separate sequence and fine-tune the BERT-base architecture on the Masked Language Modelling (MLM) objective (Devlin et al., 2019). This allows us to use the complete CaSiNo dataset in a principled way.

5.2 Experiment Design

Evaluation Metrics: We compare our methods for each strategy label on F1-score for positive class (presence of strategy label). To capture the overall performance, we report average F1 across all labels with uniform weights. Inspired by Joint Goal Accuracy from Dialog State Tracking (Kumar et al., 2020), we define another overall met-

ric called **Joint-A**, which measures the percentage of utterances for which the model predicts all the strategies correctly.

Methods: Fine-tuning the pre-trained models has achieved state-of-the-art results across many supervised tasks. Hence, our primary baseline is **BERT-FT**, which fine-tunes the BERT-base architecture for binary classification of each strategy label separately. We consider a **Majority** baseline, where the model directly outputs the majority class in the training data. We also implement a Logistic Regression model for each label separately based on a bag-of-words feature representation of the input utterance. We refer to this model as **LR-BoW**. We refer to our complete architecture presented in Figure 1 as **Full**, and consider its ablations by freezing the BERT layer (**Freeze**), removing task-specific self-attention (**No Attn**), or removing the turn position embeddings (**No Feats**). We also implement a simple over-sampling strategy where every utterance with at least one strategy is considered twice while training (referred to as **OS**). For **IDPT**, we fine-tune BERT for 20 epochs using a masking probability of 0.3. We also tried a lower masking probability of 0.15, however, in that case, the model is unable to learn anything useful on our relatively small dataset.

Training Details: Our context window considers past 3 utterances and concatenates them using an *EOS* token. The embedding dimension is 768 for the encoder and the task-specific self-attention layers, each having only one attention head. We use the turn position embeddings of 32 dimensions. We train the models with Adam optimizer with a learning rate of $5e^{-05}$ and weight decay of 0.01. We use ReLU activation for feed-forward layers, and a dropout of 0.1 to prevent overfitting. The models were trained for a maximum of 720 iterations with a batch size of 64 ($\sim$ 13 epochs). We checkpoint and evaluate the model after every 72 iterations and the best performing checkpoint on a held-out 5% validation set is used for evaluation. We provide further training details including specifics of the architecture design, computing infrastructure, and hyper-parameter tuning in Appendix F.

Results: Table 3 summarizes the results on 5-fold cross-validation. **Majority** baseline fails to recognize any of the strategies due to the data being skewed towards the negative class. It still achieves 39.4% **Joint-A**, indicating that these many utterances have none of the seven strategies present.

Model	Small-Talk F1	Self-Need F1	Other-Need F1	No-Need F1	Elicit-Pref F1	UV-Part F1	Vouch-Fair F1	Overall F1	Joint-A
Majority	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	39.6
LR-BoW	64.6	57.2	43.2	17.5	56.5	14.3	50.4	43.4	52.4
BERT-FT	81.6	72.3	76.7	16.4	80.5	20.4	61.9	58.5	64.0
Multi-task training									
Freeze	81.0	69.1	69.5	14.8	77.6	9.2	66.3	55.4	65.8
No Attn	80.7	71.9	76.8	7.5	79.0	23.2	60.6	57.1	67.8
No Feats	**82.7**	75.1	78.8	37.8	82.4	46.2	66.8	67.1	69.9
Full	**82.7**	74.4	77.9	36.4	**83.2**	44.5	**67.9**	66.7	**70.2**
+OS	82.0	**77.1**	75.6	44.2	81.9	46.4	67.3	67.8	70.1
+IDPT	82.6	74.0	**80.4**	41.2	82.8	40.8	64.0	66.6	69.5
+IDPT+OS	82.6	75.2	78.8	**46.2**	81.8	**47.3**	66.1	**68.3**	**70.2**

Table 3: Performance on strategy prediction task for 5-fold cross validation. F1 score corresponds to the positive class.
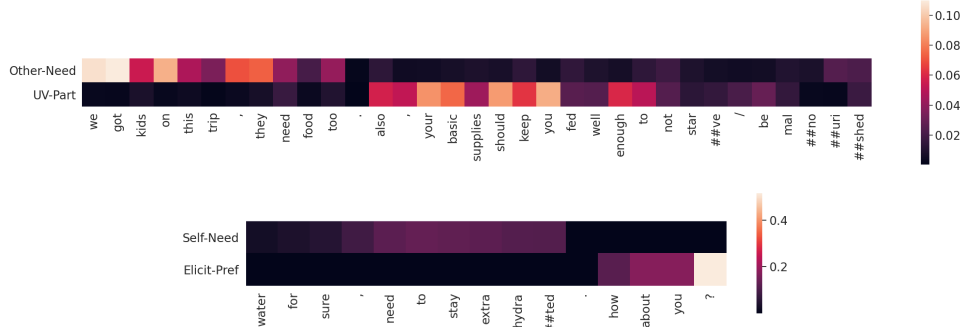

Figure 2: Visualizing task-specific self-attention layers for two examples from the test dataset for the first cv fold. The heatmap shows the attention scores for each token in the utterance for corresponding strategy labels.

Incorporating the bag-of-words features, **LR-BoW** performs much better than **Majority**. **BERT-FT** highly improves the performance on all strategies except **No-Need** and **UV-Part**, for which the dataset is the most skewed. However, our **Full** multi-tasking framework is able to tackle the imbalance in these strategies through parameter sharing between all tasks. It achieves 36.4% F1 for **No-Need** and 44.5% F1 for **UV-Part**, indicating more than 100% improvements in both the cases. The model also improves F1 scores for all other metrics, but the improvement is not that substantial. Relatively lower scores for **Freeze** and **No Attn** suggest that both fine-tuning and task-specific attention layers are essential for the performance. Turn position embeddings, however, only help for a few strategies, indicating the diverse usage of strategies in CaSiNo-Ann. Overall, we find that using over-sampling and in-domain pre-training further helps the performance, especially for **No-Need** and **UV-Part**. Although there is no clear winner among **OS** and **IDPT**, our final model, **Full+IDPT+OS**, that combines both these strategies performs the best for us, achieving an overall F1 score of 68.3% and 70.2% Joint Accuracy.

Attention Visualization: To understand if the model learns meaningful representations, we vi-

sualize the task-specific self-attention layers of the trained **Full+IDPT+OS** model. We consider two instances in Figure 2. For meaningful comparisons, the instances were picked randomly from the pool of all utterances that contain two strategies. As evident, the model is able to focus on the most relevant parts for each strategy label. For instance, in case of **Other-Need**, the scores are higher where the participant talks about their kids needing more food. The token *we* gets the most attention, which is commonly used by the participants when referring to group needs. We see similar trends in the second case as well. Remarkably, this suggests that although our annotations are at an utterance level, it might be possible to automatically retrieve the most relevant phrases for any given strategy − this requires further investigation which we aim to explore in the future.

6 Related Work

Historically, negotiations have been widely studied across multiple disciplines, in game theory (Nash Jr, 1950), understanding human behaviour (Adair et al., 2001), and building automatic negotiation agents (Beam and Segev, 1997; Baarslag et al., 2016). Most efforts focused on agent-agent interactions (Williams et al., 2012;

Lin et al., 2014; Cao et al., 2018), although there is an increasing interest in human-agent negotiations (Mell and Gratch, 2017) as well. DeVault et al. (2015) used a multi-issue bargaining design similar to ours. However, they focus on face-to-face negotiations, including speech and virtual embodied systems, which can be interesting future extensions to our current focus in chat-based dialogue systems. Other datasets looked at negotiation dialogues such as game settings (Asher et al., 2016; Lewis et al., 2017), and buyer-seller negotiations (He et al., 2018). These datasets have fueled a number of efforts on developing negotiation systems (Cheng et al., 2019; Parvaneh et al., 2019) and building a negotiation coach (Zhou et al., 2019). Our focus is on campsite negotiations, targeting a realistic and a closed-domain environment.

Several other related efforts have explored problems between task-oriented and open-domain scenarios, such as persuasion for a charity (Wang et al., 2019), anti-scam (Li et al., 2020), collecting cards in a maze (Potts, 2012), and searching for a mutual friend (He et al., 2017). Instead, we focus on rich personal negotiations, which differ from these tasks in their ultimate goal and downstream applications.

7 Conclusions and Future Work

We described the design and development of the CaSiNo dataset and the associated annotations. Our design is based on a relatable campsite scenario that promotes constrained, yet linguistically rich and personal conversations. We next plan to explore two main projects: first, extending the analysis to demographic and personality traits in the data, and second, using our insights towards the development of practical automated negotiation systems that engage in free-form dialogue and portray well-studied strategies from the prior negotiation literature. Our work fuels other tasks to advance the research in human-machine negotiations, such as predicting satisfaction and opponent perception from dialog behaviors, and building a feedback mechanism for skills training by identifying the use of pro-social versus pro-self strategies.

Finally, we note that there are many interesting extensions to our task design that make the scenario more complicated, but useful in specific realistic settings. For instance, incorporating more than two negotiating parties, and considering other modalities like facial expressions or embodied agents. In some realistic settings, the individual preferences may change during the negotiation and our setup assumes a fixed set of preferences throughout. Further, in complex settings, it may be possible to break down an individual item and claim sub-parts, such as negotiating for who gets an orange, but one party ends up taking the husk and the other takes the pulp for their own purposes. This is again not considered in our work and opens up exciting avenues for future work.

8 Broader Impact and Ethical Considerations

8.1 Data Collection

Our study was approved by our Institutional Review Board (IRB). Each participant signed an Informed Consent document at the beginning of the study which covered the purpose of the study, warned about potential discomfort, and noted the collection of data and its later use. Further, the participants were informed that they can withdraw at any time. They were also instructed to not use any offensive or discriminative language. The compensation was determined in accordance with the fairness rules defined by our IRB approval process. Additionally, we release the anonymized version of the data for future work by the research community. All personally identifiable information such as MTurk Ids or HIT Ids was removed before releasing the data. Lastly, any mention of the demographics or the psychological personality of the participants is based on self-identified information in our pre-survey and standard procedures of collecting personality metrics in the literature.

8.2 Automatic Negotiation Systems

Students entering the modern workforce must have a number of interpersonal skills that are crucial across a wide range of jobs. One of the key interpersonal skills needed to address conflicts and work well with others is the ability to negotiate. Unfortunately, research shows that the average human is bad at negotiating. This can adversely impact work opportunities (Babcock and Laschever, 2009), legal settlements (Eisenberg and Lanvers, 2009), and cross-cultural border peace (Wunderle, 2007). The typical way to teach negotiation skills to students is by in-class simulations, which are expensive. Automated systems can dramatically reduce the costs of, and increase access to, negotiation training. Systems developed on CaSiNo would be useful in this context. Further, the techniques

developed find use-cases for advancing conversational AI and imparting the negotiation skills to existing AI assistants, making them more aware of our preferences and requirements. One such prototype is Google Duplex (Leviathan and Matias, 2018), where the AI system engages in a simple form of negotiation to book a haircut appointment over the phone.

How humans negotiate has been actively studied for decades in Economics, Psychology, and Affective Computing (Carnevale and Pruitt, 1992). With this huge progress in our understanding of human-human negotiations, ethics has also been a well-studied topic in the literature (Lewicki et al., 2016). Primary concerns include the acts of emotion manipulation, deception, bias, and misrepresentation. Naturally, these ethical concerns may creep into the automated systems, trained on a human-human negotiation dataset.

To mitigate these ethical impacts, we recommend that standard guidelines for deploying conversational AI assistants should be followed. It is essential to maintain transparency about the identity of the system. Ethical principles must be in place before the deployment of such systems with a regular update cycle. Our camping scenario is quite relatable to anyone who negotiates with the system, hence, it is important to be upfront about the potential behaviors of the deployed system. We recommend continuous monitoring by keeping humans in the loop, ensuring that the system is neither offensive nor discriminative. Further, it should be made easy for the users negotiating with the system to directly contact the team behind the deployment. Finally, any data which is collected during the deployment phase should be informed to the users and its future purpose should be properly laid out.

Acknowledgments

References

Wendi L Adair, Tetsushi Okumura, and Jeanne M Brett. 2001. Negotiation behavior when cultures collide: the united states and japan. *Journal of Applied Psychology*, 86(3):371.

J Stacy Adams. 1976. The structure and dynamics of behavior in organizational boundary roles. *Handbook of industrial and organizational psychology*, 1175:1199.

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the stac corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727.

Reyhan Aydoğan, Tim Baarslag, Katsuhide Fujita, Johnathan Mell, Jonathan Gratch, Dave de Jonge, Yasser Mohammad, Shinji Nakadai, Satoshi Morinaga, Hirotaka Osawa, et al. 2020. Challenges and main results of the automated negotiating agents competition (anac) 2019. In *Multi-Agent Systems and Agreement Technologies*, pages 366–381. Springer.

Tim Baarslag, Mark JC Hendrikx, Koen V Hindriks, and Catholijn M Jonker. 2016. A survey of opponent modeling techniques in automated negotiation. In *15th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2016*, pages 575–576. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS).

Linda Babcock and Sara Laschever. 2009. *Women don't ask: Negotiation and the gender divide*. Princeton University Press.

Carrie Beam and Arie Segev. 1997. Automated negotiations: A survey of the state of the art. *Wirtschaftsinformatik*, 39(3):263–268.

Sandy Bogaert, Christophe Boone, and Carolyn Declerck. 2008. Social value orientation and cooperation in social dilemmas: A review and conceptual model. *British Journal of Social Psychology*, 47(3):453–480.

Kris Cao, Angeliki Lazaridou, Marc Lanctot, Joel Z Leibo, Karl Tuyls, and Stephen Clark. 2018. Emergent communication through negotiation. In *International Conference on Learning Representations*.

Peter J Carnevale and Dean G Pruitt. 1992. Negotiation and mediation. *Annual review of psychology*, 43(1):531–582.

Minhao Cheng, Wei Wei, and Cho-Jui Hsieh. 2019. Evaluating and enhancing the robustness of dialogue systems: A case study on a negotiation agent. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3325–3335.

Stephen W Clopton. 1984. Seller and buying firm factors affecting industrial buyers' negotiation behavior and outcomes. *Journal of Marketing Research*, 21(1):39–53.

Rachel G Curtis, Tim D Windsor, and Andrea Soubelet. 2015. The relationship between big-5 personality traits and cognitive ability in older adults–a review. *Aging, Neuropsychology, and Cognition*, 22(1):42–71.

David DeVault, Johnathan Mell, and Jonathan Gratch. 2015. Toward natural turn-taking in a virtual human negotiation agent. In *AAAI Spring Symposia*. Citeseer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Theodore Eisenberg and Charlotte Lanvers. 2009. What is the settlement rate and why should we care? *Journal of Empirical Legal Studies*, 6(1):111–146.

Chaim Fershtman. 1990. The importance of the agenda in bargaining. *Games and Economic Behavior*, 2(3):224–238.

Lewis R Goldberg. 1990. An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*, pages 1766–1776. Association for Computational Linguistics (ACL).

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343.

Emmanuel Johnson, Gale Lucas, Peter Kim, and Jonathan Gratch. 2019. Intelligent tutoring system for negotiation skills training. In *International Conference on Artificial Intelligence in Education*, pages 122–127. Springer.

Harold H Kelley. 1996. *A classroom study of the dilemmas in interpersonal negotiations*. Berkeley Institute of International Studies.

Olga M Klimecki. 2019. The role of empathy and compassion in conflict resolution. *Emotion Review*, 11(4):310–325.

Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Adarsh Kumar, Peter Ku, Anuj Goyal, Angeliki Metallinou, and Dilek Hakkani-Tur. 2020. Ma-dst: Multi-attention-based scalable dialog state tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8107–8114.

Yaniv Leviathan and Yossi Matias. 2018. Google duplex: An ai system for accomplishing real-world tasks over the phone. *URL https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html*, 3.

Roy J Lewicki, Bruce Barry, and David M Saunders. 2016. *Essentials of negotiation*. McGraw-Hill.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In *EMNLP*.

Yu Li, Kun Qian, Weiyan Shi, and Zhou Yu. 2020. End-to-end trainable non-collaborative dialog system. In *AAAI*, pages 8293–8302.

Raz Lin, Sarit Kraus, Tim Baarslag, Dmytro Tykhonov, Koen Hindriks, and Catholijn M. Jonker. 2014. Genius: An integrated environment for supporting the design of generic automated negotiators. *Computational Intelligence*, 30(1):48–70.

Johnathan Mell and Jonathan Gratch. 2016. Iago: interactive arbitration guide online. In *AAMAS*, pages 1510–1512.

Johnathan Mell and Jonathan Gratch. 2017. Grumpy & pinocchio: answering human-agent negotiation questions through realistic agent design. In *Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems*, pages 401–409. International Foundation for Autonomous Agents and Multiagent Systems.

Johnathan Mell, Jonathan Gratch, Reyhan Aydoğan, Tim Baarslag, and Catholijn M Jonker. 2019. The likeability-success tradeoff: Results of the 2 nd

annual human-agent automated negotiating agents competition. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. Parlai: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84.

Shannon Monahan, Emmanuel Johnson, Gale Lucas, James Finch, and Jonathan Gratch. 2018. Autonomous agent that provides automated feedback improves negotiation skills. In *International Conference on Artificial Intelligence in Education*, pages 225–229. Springer.

Janice Nadler. 2003. Rapport in negotiation and conflict resolution. *Marq. L. Rev.*, 87:875.

John F Nash Jr. 1950. The bargaining problem. *Econometrica: Journal of the Econometric Society*, pages 155–162.

Amin Parvaneh, Ehsan Abbasnejad, Qi Wu, and Javen Shi. 2019. Show, price and negotiate: A hierarchical attention recurrent visual negotiator. *arXiv preprint arXiv:1905.03721*.

Christopher Potts. 2012. Goal-driven answers in the cards dialogue corpus. In *Proceedings of the 30th west coast conference on formal linguistics*, pages 1–20. Cascadilla Proceedings Project.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.

Paul AM Van Lange, Ellen De Bruin, Wilma Otten, and Jeffrey A Joireman. 1997. Development of prosocial, individualistic, and competitive orientations: theory and preliminary evidence. *Journal of personality and social psychology*, 73(4):733.

Paul AM Van Lange, David De Cremer, Eric Van Dijk, and Mark van Vugt. 2007. 23. self-interest and beyond: basic processes of social interaction. In *Social psychology: handbook of basic principles*, pages 540–564.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649.

Colin R Williams, Valentin Robu, Enrico H Gerding, and Nicholas R Jennings. 2012. Iamhaggler: A negotiation agent for complex environments. In *New Trends in Agent-based Complex Automated Negotiations*, pages 151–158. Springer.

William Wunderle. 2007. How to negotiate in the middle east. *Military review*, 87(2):33.

Toshio Yamagishi, Yoshie Matsumoto, Toko Kiyonari, Haruto Takagishi, Yang Li, Ryota Kanai, and Masamichi Sakagami. 2017. Response time in economic games reflects different types of decision conflict for prosocial and proself individuals. *Proceedings of the National Academy of Sciences*, 114(24):6394–6399.

Yiheng Zhou, He He, Alan W Black, and Yulia Tsvetkov. 2019. A dynamic strategy coach for effective negotiation. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 367–378.

A Pre-Survey

After an internal pilot with 9 participants, the entire CaSiNo dataset was collected on Amazon Mechanical Turk over a period of a month. In total, 846 subjects took part in our data collection study. The statistics presented in this section are based on self-identified demographical attributes and standard ways of collecting personality traits from the literature. We had a highly diverse participant pool, representing different age groups, gender, ethnic backgrounds and education levels. The mean *Age* among our participants is 36.97 with a standard deviation of 10.81. One participant was removed from this computation since the age entered was 3, which we believed to be in error. Among the participants, 472 identified themselves as *Female*, 372 were *Male*, and *2* belonged to *Other* category. While most of the participants were *White American* (625 in count), our study also involved a mix of *Asian American*, *Black or African American*, *Hispanic or Latino*, and *Multi-Racial* groups, among others. Most common *highest level of education* was found to be a 4-year Bachelor degree (346 participants), although the complete pool represents a mixture of Master and PhD degree holders, 2-year and 4-year college graduates without degrees, and high school graduates, among others.

For the personality traits, 364 participants were classified as Proself, 463 as Prosocial, and 19 were unclassified based on their Social Value Orientation[2]. The mean scores for the Big-5 personality traits were found to be as follows: Agreeableness: 5.27, Conscientiousness: 5.6, Emotional Stability: 4.91, Extraversion: 3.69, Openness to Experiences: 5.04. We use the Ten-Item Personality Inventory (TIPI)[3] to compute these attributes, where each of them takes a value between 1 and 7.

B Preparation Phase

We present the scenario description seen by the participants in Table 4. Several arguments that the participants come up with are presented in Table 5.

> *Imagine that you are on a camping trip! Woohoo! Apart from some basic amount of supplies which are provided to everyone, you can collect some additional food packages, water bottles and firewood, to make your camping trip even better. Since these are limited in quantity, you will have to split these additional packages with your campsite neighbor!*
>
> *Each of these items will be of either High, Medium or Low priority for you. Each of them only has an available quantity of 3. You will negotiate with another MTurker by chatting in English, using reasons from your personal experiences to justify why you need additional packages apart from the basic supplies. Try hard to get as many items as you can!*

Table 4: The camping scenario description as seen by the participants in our data collection.

C Data Post-processing steps

We list the data post-processing and filtering steps below:

1. **Removal of incomplete dialogues**: During the data collection, many negotiation sessions could not be completed due to one of the participants' disconnecting in the middle. Any dialogue for which we had missing data, including pre-survey and post-survey responses for both the participants, was removed from the final dataset.

2. **Removal of bad quality dialogues**: We also removed dialogues where we observed a lack of effort or an irrelevant dialogue between the participants. We removed dialogues where the participants used very short utterances or failed to answer the dummy questions about their own preferences correctly, suggesting a lack of effort. Further, we removed the instances where the participants talked about the MTurk task itself, rather than the negotiation. These cases were identified based on a list of keywords: {'mturk', 'amt', 'turns', 'messages', 'amazon', '10'}. In a few cases, it was possible to retain the complete dialogue structure by just removing a few utterances. Hence, in these cases, we only removed the irrelevant utterances, while retaining the rest of the dialogue and the associated metadata.

3. **Tackling inappropriate language use**: Rarely, some participants also used inappropriate language in their utterances. These dialogues were identified using the lexicon

Category	Item type		
	Food	**Water**	**Firewood**
Personal Care	because I'm normally eat more because of my big size	I have to take a lot of medicine so hydration is very important	I have arthritis and being sure I am warm is important for my comfort.
Recreational	Need many snacks throughout the day for energy to hike	I am a very active camper. I like to hike when I camp and I once ran out of water during a strenuous hike.	I like having campfires so I need all the firewood.
Group Needs	I have two teenage boys who require a lot of food, especially when expending so much energy with all the activities of camping.	I need more water because I have more people to keep hydrated and do not have enough.	I need more firewood due to having several people join on the trip and needing a bigger fire overall.
Emergency	Some could have been damaged during the trip. I would need more.	our car overheated we had to use the water	It may get cold and firewood can be hard to come by at certain campsites.

Table 5: Example arguments that the participants come up for their individual requirements during the preparation phase. The categories defined are not exhaustive.

D Participant Feedback

Role-playing has been a key technique to teach negotiation skills in classroom settings. One of the key application areas for automated negotiation systems is to augment such exercises by allowing the human participants to negotiate with an AI and practice their social skills. To maximize the utility of the system developed using our dataset, we choose the camping scenario, which we expected to be easily relatable for our participants and also for any individual who negotiates with a system developed on our dataset. This is essential to ensure that the collected dialogues are engaging, interesting, and capture the rich personal context of the individuals, albeit in a closed-domain setting. One way to judge whether the participants are able to relate to the scenario is via their feedback after the study. With this in mind, we used a feedback column in the Post-survey and asked several questions to the participants throughout the data collection process. These questions included: 1) How was your overall experience? 2) Were you able to see yourself in the 'role' and follow best practices?, 3) Could you relate to camping?, and 4) How helpful was the preparation phase?

Based on manual inspection, we observed an overall positive feedback for all the above questions. Most of the participants were able to easily relate to camping. They frequently pointed out that the experience was 'fun', 'interesting', and 'nice'. Many saw this as an opportunity to talk to someone during these tough times of the pandemic. Several cherry-picked feedback responses which indicate that the participants enjoyed the task as a whole and were in fact able to connect well and engage in the negotiation, have been provided in Table 6.

E Correlational Analysis

The analysis discussed in the paper is presented in Tables 7, 8, 9, and 10.

F Strategy Prediction

F.1 Architecture

We provide some more details on the strategy prediction multi-task architecture in this section. The self-attention layer is itself represented using the BERT encoder architecture, but with a single transformer layer and just one attention head. After the self-attention layer, we first extract the 768 dimensional representation for the [CLS] token. This is passed through a feed-forward network, which converts it to 128 dimensions. The feature embedding is also converted to a 128 dimensional vector using a feed-forward network. Both the above embeddings are then combined using an element-wise summation, which further passes through two feed-forward layers with hidden dimensions of 64 and 1, and a sigmoid layer to finally output the probability for each annotation strategy.

of English swear words on Wikipedia[4]. All these dialogues were also removed from the final dataset.

I could do this all day	
I am camping right now!	
My partner had better reasons for needing the firewood	
I enjoyed talking about camping, I haven't been in a while. It reminded me of all of the things that I used to do.	
The best thing I did was ask him what his preferences were. He had no interest in firewood which was my highest priority.	

Table 6: A few positive feedback responses which we obtained from the participants during the collection of the CaSiNo dataset.

	Points-Scored	Satisfaction	Opp-Likeness
Points-Scored	1	.376**	.276**
Satisfaction	.376**	1	.702**
Opp-Likeness	.276**	.702**	1
P.Points-Scored	−.092**	.105**	.132**
P.Satisfaction	.105**	.180**	.244**
P.Opp-Likeness	.132**	.244**	.344**

Table 7: Pearson Correlation Coefficients (r) between the outcome variables. Variables with **P.** prefix denote the corresponding attributes of the negotiation partner of an individual. These correlations have been computed on the entire CaSiNo dataset. * denotes significance with $p < 0.05$ (2-tailed). ** denotes significance with $p < 0.01$ (2-tailed).

F.2 Computing Infrastructure

All experiments were performed on a single Nvidia Tesla V100 GPU. The training takes two hours to complete for a single model on all the cross-validation folds.

F.3 Training Details

To search for the best hyperparameters, we use a combination of randomized and manual search for the **Full** model. For each cross fold, 5% of the training data was kept aside for validation. The metric for choosing the best hyper-parameters is the mean F1 score for the positive class on the validation dataset. The mean is over all the labels and over 5 cross-validation folds.

We vary the learning rate in $\{3e^{-5}, 4e^{-5}, 5e^{-5}\}$, weight decay in $\{0.0, 0.01, 0.001\}$ and dropout in $\{0.0, 0.1, 0.2, 0.3\}$. The rest of the hyper-parameters were fixed based on the available computational and space resources. We report the best performing hyper-parameters in the main paper, which were used for all the experiments. We report the performance on the validation set corresponding to the chosen hyper-parameters and the number of trainable parameters in Table 11.

G Screenshots from the data collection interface

To provide more clarity on the data collection procedure, we provide several screenshots from our interface in Figures 3, 4, 5, and 6. We design the

pre-survey using the Qualtrics platform[5]. The rest of the data collection is based on the ParlAI framework (Miller et al., 2017).

	Joint Points
Integrative potential	.425***

Table 8: Pearson Correlation Coefficient (r) between integrative potential and the joint negotiation performance. *** denotes significance with $p < 0.001$.

	Joint Points	Points-Scored	Satisfaction	Opp-Likeness	P.Points-Scored	P.Satisfaction	P.Opp-Likeness
	Prosocial Generic						
Small-Talk	−.022	−.002	.086*	.115**	−.025	.068	.127**
	Prosocial About Preferences						
No-Need	−.003	−.066	.035	.023	.063	.083*	.089*
Elicit-Pref	.053	.055	.058	.015	.010	.022	.055
	Proself Generic						
UV-Part	−.037	.008	−.051	−.112**	−.054	−.131**	−.151**
Vouch-Fairness	−.140**	−.084*	−.159**	−.196**	−.090*	−.185**	−.180**
	Proself About Preferences						
Self-Need	−.003	.022	−.061	−.065	−.026	−.091*	−.086*
Other-Need	−.176**	−.045	−.101**	−.118**	−.174**	−.160**	−.113**

Table 9: Pearson Correlation Coefficients (r) for strategy annotation counts with the outcome variables. Variables with **P.** prefix denote the corresponding attributes of the negotiation partner of an individual. These correlations have been computed on the annotated subset of the CaSiNo dataset. * denotes significance with $p < 0.05$ (2-tailed). ** denotes significance with $p < 0.01$ (2-tailed).

	P.Small-Talk	P.Self-Need	P.Other-Need	P.No-Need	P.Elicit-Pref	P.UV-Part	P.Vouch-Fair
Small-Talk	.769**	−.033	.021	.063	−.059	−.012	−.180**
Self-Need	−.033	.355**	.103**	.115**	−.007	.235**	−.088*
Other-Need	.021	.103**	.339**	.002	−.067	.159**	−.015
No-Need	.063	.115**	.002	.258**	.097**	.064	−.116**
Elicit-Pref	−.059	−.007	−.067	.097**	.168**	−.097**	−.102**
UV-Part	−.012	.235**	.159**	.064	−.097**	.268**	.064
Vouch-Fair	−.180**	−.088*	−.015	−.116**	−.102**	.064	.287**

Table 10: Pearson Correlation Coefficients (r) between strategy annotation counts. Variables with **P.** prefix denote the corresponding attributes of the negotiation partner of an individual. These correlations have been computed on the annotated subset of the CaSiNo dataset. * denotes significance with $p < 0.05$ (2-tailed). ** denotes significance with $p < 0.01$ (2-tailed).

Model	Overall Validation F1	Trainable Parameters
Majority	0.0	0
LR-BoW	49.6	2646.2 (27.2)
BERT-FT	69.9	109, 590, 529
	Multi-task training	
Freeze	62.3	221, 361, 031
No Attn	66.6	110, 235, 271
No Feats	77.6	330, 840, 583
Full	78.1	330, 844, 807
+OS	77.9	330, 844, 807
+IDPT	79.6	330, 844, 807
+IDPT+OS	79.6	330, 844, 807

Table 11: Training details for the strategy prediction task. The Overall F1 scores are for the positive class. For **LR-BoW**, the exact number of features varies slightly based on the CV split. Hence, we report Mean (Std) across the five splits.

Negotiate a deal with another user! Earn upto $5 instantly for one HIT.

You will first be given a quick tutorial on how to negotiate. Then, you will negotiate with another MTurker to arrive at an agreement. You will earn $0.30 base pay and $1.70 on completion (when none of the partners disconnects mid-way). Additionally, you can earn upto $3 in bonus for one HIT, depending on performance and quality. The task is about 20 minutes. Turkers with too many disconnects will be blocked for similar HITs in the future.

Figure 3: Screenshots from the data collection interface: Task Preview. This is a brief task description which the MTurkers see before signing up for our data collection task.
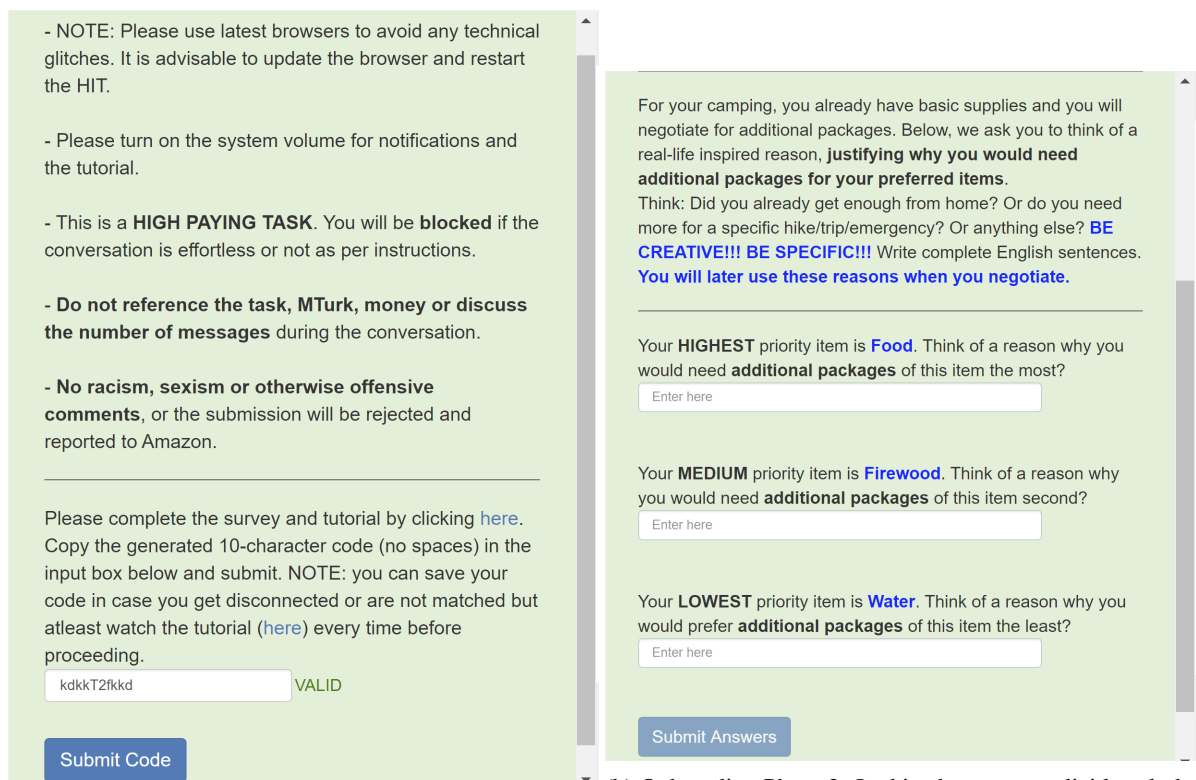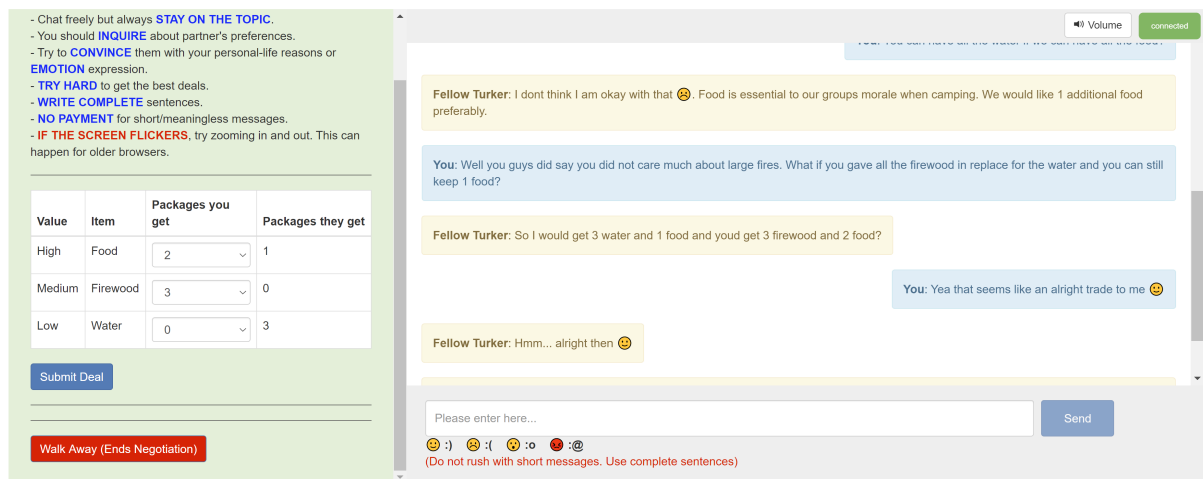
(a) Onboarding Phase 1: The first step takes the participant to Qualtrics which collects the demographics, introduces the camping scenario and gives a tutorial on negotiation best practices.

(b) Onboarding Phase 2: In this phase, we explicitly ask the participants to come up with arguments from their past experiences, which justify their preferences. The preference order is randomly assigned by us. This provides a personal context around the negotiation for each participant.

Figure 4: Screenshots from the data collection interface: Participant On-boarding.

(a) Chat Interface: The right portion allows two participants to negotiate in English using alternating messages. They also have the option to use emoticons. Once they come to an agreement, one of the participant must enter the exact deal on the left.
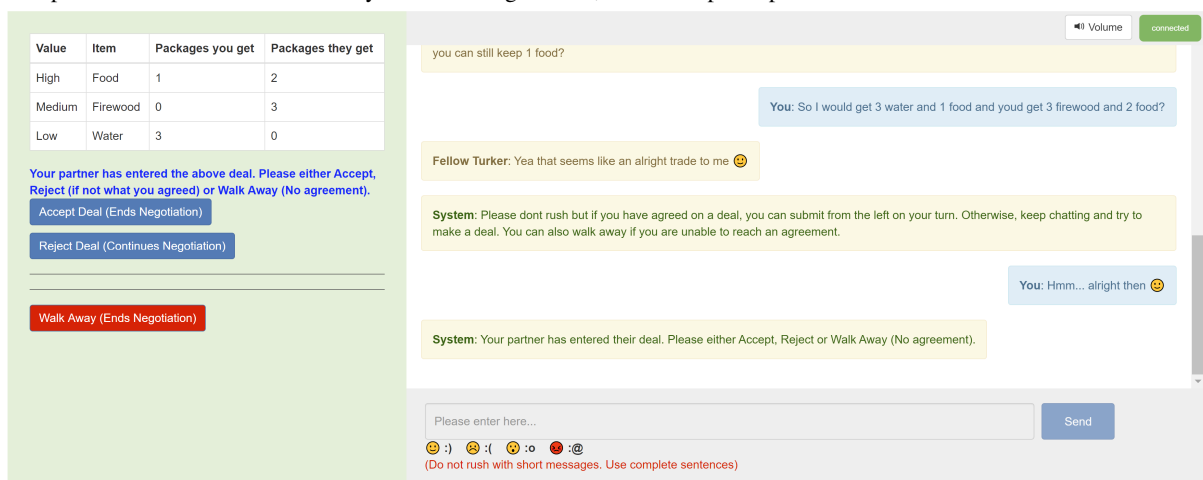
(b) Response to the Deal: When one of the participants enters the deal, the other gets an option to either accept, reject, or walk away from the deal. In the CaSiNO dataset, a participant walks away in 36 dialogues.

Figure 5: Screenshots from the data collection interface: Chat Interface.

Figure 6: Screenshots from the data collection interface: Post-Survey. Once the deal is accepted (or someone walks away), both the participants are asked to fill in the post-survey having the above questions. The figure contains dummy responses.