Dialogue Act-based Breakdown Detection in Negotiation Dialogues

Atsuki Yamaguchi[♠]*, **Kosui Iwasa**[♣] **and Katsuhide Fujita**[♢]
[♠]Research and Development Group, Hitachi, Ltd.
[♣]Graduate School of Engineering, Tokyo University of Agriculture and Technology
[♢]Institute of Engineering, Tokyo University of Agriculture and Technology
[♠]atsuki.yamaguchi1@gmail.com
[♣]contact@iwasakosui.com
[♢]katfuji@cc.tuat.ac.jp

Abstract

Thanks to the success of goal-oriented negotiation dialogue systems, studies of negotiation dialogue have gained momentum in terms of both human-human negotiation support and dialogue systems. However, the field suffers from a paucity of available negotiation corpora, which hinders further development and makes it difficult to test new methodologies in novel negotiation settings. Here, we share a human-human negotiation dialogue dataset in a job interview scenario that features increased complexities in terms of the number of possible solutions and a utility function. We test the proposed corpus using a breakdown detection task for human-human negotiation support. We also introduce a dialogue act-based breakdown detection method, focusing on dialogue flow that is applicable to various corpora. Our results show that our proposed method features comparable detection performance to text-based approaches in existing corpora and better results in the proposed dataset.

1 Introduction

Negotiation is an essential task involved in our daily life. In negotiation, people work to maximize their profits by bargaining; however, negotiation sometimes breaks down due to conflicts between people's competing interests. To help them to reach rational agreement, previous studies of multiagent systems have proposed the use of negotiating agents (Lin and Kraus, 2010; Jonker et al., 2017; Baarslag et al., 2013a). Recently, several studies have succeeded in modeling a negotiating agent in natural language that can control both text generation and reasoning in the context of goal-oriented dialogue systems, and such agents have produced better performance than human players in some cases (Lewis et al., 2017; He et al., 2018; Cheng et al., 2019). Further, support for human-human negotiation in natural language has also been tackled, involving negotiation corpora developed for goal-oriented dialogue systems, such as a Nash bargaining solution estimation (Iwasa and Fujita, 2018), real-time negotiation coaching (Zhou et al., 2019), and negotiation breakdown detection (Yamaguchi and Fujita, 2020).

Although they have recently attracted additional attention, there are only few negotiation corpora, as the most recent follow-up studies (Iwasa and Fujita, 2018; Cheng et al., 2019; Zhou et al., 2019; Yamaguchi and Fujita, 2020) have only utilized either the DEALORNODEAL (DN) (Lewis et al., 2017), CRAIGSLISTBARGAIN (CB) (He et al., 2018) datasets or both. Moreover, most existing corpora have simplified negotiation settings; for example, the DN dataset handles the negotiation of item division between humans with 22.5 possible solutions per dialogue and uses a standard linear additive utility function (Keeney and Raiffa, 1993; Raiffa et al., 2002) for scoring. The CB dataset is only concerned with price negotiation on a listed product between two human negotiators. These settings might make it easy for a machine learning (ML) model to reach optimal solution or fulfill its goal. Finally, some existing corpora (Konovalov et al., 2016; Petukhova et al., 2016; Asher et al., 2016) other than the DN and CB datasets have far smaller samples (scenarios), which makes it challenging to use them for goal-oriented dialogue systems or end-to-end human-human negotiation support. All of these factors inhibit further development in the field and its future applicability to real-world problems. Furthermore, no effective breakdown detection method for negotiation dialogues has been proposed. Negotiation features certain unique characteristics relative to other dialogues, such as offering proposals, accepting them,

*This work was conducted when the first author was a master's student at the University of Sheffield, UK.

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 745–757
April 19 - 23, 2021.

and making counter-offers (Thompson et al., 2010; Traum et al., 2008). If the breakdown detection method can incorporate these characteristics, the quality of breakdown detection will be improved.

This study proposes a new negotiation corpus in a job interview setting with increased complexities relative to a range of solutions and a utility function. We enact a breakdown detection task (Yamaguchi and Fujita, 2020) across three negotiation datasets including a proposed one with a novel dialogue act-based approach that can focus on dialogue flow. This task can support human-human negotiation by alerting negotiators to potential breakdowns, which prevents the loss of time and negotiator utility. We highlight the following contributions:

1. We develop a new English negotiation corpus for a job interview setting, consisting of 2639 crowd-sourced dialogues (Section 3).

2. We propose a novel breakdown detection method that employs dialogue act-based features and a gated recurrent unit (GRU) (Chung et al., 2014)-based model (Section 5).

3. We demonstrate that the proposed method exhibits results that are comparable to models with text-based features in the existing corpora and outperforms them in the proposed corpus, which has a far smaller breakdown ratio (Section 7).

4. We conduct ablation studies and error analyses to examine how our proposed features works on a GRU-based model (Section 7).

2 Related Work

Automated Negotiation in Multiagent Systems Automated negotiation is a field of research, in which computers negotiate with each other and try to seek appropriate agreement without human intervention (Baarslag et al., 2013a). Typical applications include supply chain management (Wang et al., 2009) and smart grids (Ketter et al., 2013). As automated negotiation has gained momentum, the International Automated Negotiating Agents Competition (ANAC) (Baarslag et al., 2015; Jonker et al., 2017) has been being held annually since 2010. This event encourages the development of state-of-the-art negotiating strategies for automated negotiating agents in both agent-agent and human-agent (Mell et al., 2018) negotiations. The major difference between automated negotiation and ours is that the former supports negotiation by letting the agents negotiate instead of humans, whereas the latter seeks to support human-human negotiation in natural language only by providing feedback to negotiators with ML models.

NLP for Human-human Negotiation Support Automated negotiation has gained a great deal attention, but there have been only a few studies conducted on support for human-human negotiation in natural language: Iwasa and Fujita (2018) have proposed a GRU-based model to suggest a draft agreement that maximizes the sum of utilities based on the estimated weights of all items in the DN dataset. Zhou et al. (2019) proposed a dynamic negotiation coaching method in the setting of CB dataset that provides useful recommendations to sellers, resulting in increased profits. Our work is a follow-up study to Yamaguchi and Fujita (2020), who demonstrated that neural-network (NN)-based models trained with text-based features could capture signs of breakdowns in DN and CB datasets. Here, we show that text-based methods cannot detect breakdowns in the proposed corpus relative to our dialogue act-based approach.

Negotiation Dialogue Systems Previous efforts on building negotiation dialogue systems initially focused on modeling strategic aspects (Cuayáhuitl et al., 2015; Keizer et al., 2017; Petukhova et al., 2017), to construct an agent that could outperform human players by controlling a discrete action space. By contrast, Lewis et al. (2017) and He et al. (2018) have recently tried to simultaneously handle both text generation and reasoning by employing end-to-end neural negotiating models; moreover, Cheng et al. (2019) proposed adversarial training to improve the robustness of goal-oriented models. Although our main scope is supporting human-human negotiation, our corpus can also be used for goal-oriented dialogue systems (Lewis et al., 2017; He et al., 2018; Cheng et al., 2019) as its fundamental design is drawn from the DN dataset.

Negotiation Dialogue Datasets Several negotiation dialogue corpora, along with the DN and CB datasets, have been proposed to model strategic dialogue. Konovalov et al. (2016) built a bilateral negotiation corpus between a human and an agent in relation to terms of employment. Petukhova et al. (2016) created a corpus in which each negotiator acts as either a city councilor or a small business owner and debates new anti-smoking regulations.

Issue	Option
Salary	$20 to $50 per hour (integer)
Weekly day off	2 days to 5 days (integer)
⋆Position	{Engineer, Designer, Manager, Sales}
⋆Company	{Google, Apple, Facebook, Amazon}
Workplace	{Tokyo, Seoul, Beijing, Sydney}

Table 1: List of issues and options in the JI dataset: ⋆ denotes that there is an interdependent relationship between issues.

Asher et al. (2016) developed a multilateral negotiation dialogue corpus in the *Settlers of Catan* game. Our corpus and that of Konovalov et al. (2016) are similar to each other, in that both handle a job contract scenario. However, three main differences appear between the two: (1) The former handles human-human negotiation, whereas the latter deals with human-agent negotiation. (2) The former considers 11.5 times more possible solutions per dialogue than the latter. (3) The former has 2639 dialogues, and the latter has 105.

Dialogue Breakdown Detection Challenge The recently held Dialogue Breakdown Detection Challenge (DBDC) (Higashinaka et al., 2016; Hori et al., 2019) was intended to improve the coherency of a dialogue system. Given a dialogue history between a human and a system, the task is to evaluate whether a certain system response is valid. By contrast, our study focuses on predicting negotiation outcomes based on human-human negotiation to avoid negotiation breakdowns; that is, our task is different from the DBDC.

3 Job Interview Negotiation Dataset

3.1 Overview

The JOBINTERVIEW (JI) dataset is an instance of multi-issue multi-option negotiation, which includes the preferences of the negotiators, a dialogue history, proposed offers, and a settled agreement in a job interview setting. The negotiators conduct a conversation in English in the roles of recruiter or applicant and negotiate regarding the issues listed in Table 1 to maximize their scores. A dialogue sample from the JI dataset is shown in Table 2[1].

[1]Our dataset is publicly available on GitHub: https://github.com/gucci-j/negotiation-breakdown-detection, and details on the negotiation interface and procedures are given in Appendix A.

Utterance	Dialogue Act
Recruiter - Hello	<greet>
Worker - Hi	<greet>
Recruiter - I have a position open as an engineer at google. Are you interested?	<inquire>
Worker - Yes.	<agree>
Recruiter - The position is in tokyo. It pays $35/hr and it is 4 days a week. Is this acceptable?	<propose><inquire>
Worker - Salary is too low.	<disagree>
Recruiter - OK. We could bump it to $40/hr. Is this OK?	<agree><propose><inquire>
Worker - How about I work in Beijing?	<inquire>
Recruiter - Beijing is open also.	<inform>
Worker - 5 days/wk.	<propose>
Recruiter - I cannot do 5 days a week.	<disagree><propose>
Worker - 4 days/wk and $47/hr?.	<propose><inquire>
Recruiter - OK.	<agree>

Table 2: Sample dialogue between two negotiators in the JI dataset with extracted dialogue acts.

3.2 Mathematical Design

To make the negotiation competitive, we define each negotiator's preferences, and a scoring function, as in Lewis et al. (2017). In addition, we consider the interdependency (Kardan and Janzadeh, 2008; Alam et al., 2013) between a pair of issues such that the negotiators cannot easily reach an optimal agreement (Ito et al., 2006), leading them to seek a compromise solution through dialogue.

Preferences The importance of each issue and option, and bias assignment in representing interdependency between specific issues are defined as follows. Two negotiators $\boldsymbol{A} = \{a_1, a_2\}$ participate in a negotiation over the set of independent issues $\boldsymbol{I}$ and of issues $\boldsymbol{J}$ with an interdependent relationship. An issue $i \in \boldsymbol{I}$ is assigned a weight (importance) $w_i^{a_k} \in [0.1, 0.6], \sum_{i \in \boldsymbol{I}} w_i^{a_k} = 1$ per negotiator a_k with $k = 1$ or 2. An option for i, $o^i \in \boldsymbol{O}^i$, is assigned a weight $w_{o^i}^{a_k} \in [0, 1]$. While an issue included in a set of specific issues with an interdependent relationship $(j_{\text{from}}, j_{\text{to}}) \in \boldsymbol{J}^2$ has its own weight per a_k, only an option of j_{to} has a bias for that of j_{from} and j_{to}; that is, $o^{j_{\text{from}}}$ does not have a bias. The bias $b^{w_{o^{j_{\text{to}}}}}_{(o^{j_{\text{to}}}, o^{j_{\text{from}}})} \in [0, 0.5]$ represents an increase of importance for $o^{j_{\text{to}}}$ in a

[2]In our implementation, j_{from} is equivalent to "position," and j_{to} corresponds to "company."

	JI	DN	CB
# of dialogues	2,639	6,251	6,682
Avg turns per dialogue	12.7	4.97	7.53
Avg words per turn	6.12	8.56	13.60
Vocab size	4,476	2,631	12,139
Agreed [%]	92.9	76.2	74.9
PO solutions [%]	13.4	75.0	
PO bids for all bids [%]	0.98	18.0	
# of all bids per dialogue	9,920	22.5	
Avg score	6.4 / 10	5.7 / 10	

Table 3: Quantitative comparison of the three negotiation datasets: "PO" stands for Pareto optimal.

particular pair of options $(o^{j_{\text{to}}}, o^{j_{\text{from}}})$. Note that each weight and bias is initialized using uniform random numbers within a predefined range.

Scoring Function We define a scoring (utility) function to calculate a negotiation score. The weight of option $w_{o^{j_{\text{to}}}}$ is normalized after considering bias. More specifically, when an option $o^{j_{\text{from}}}$ is in a draft agreement, the normalized weight of the option $w'_{o^{j_{\text{to}}}}$ is calculated using min-max normalization of $w_{o^{j_{\text{to}}}} + b^{w_{o^{j_{\text{to}}}}}_{(o^{j_{\text{to}}}, o^{j_{\text{from}}})}$ over $\boldsymbol{O}^{j_{\text{to}}}$. Thus, the scoring function is defined as follows:

$$U_{a_k}(s) = \sum_{i \in \boldsymbol{I}} w_i^{a_k} w_{o_s^i}^{a_k} + \sum_{(j_{\text{from}}, j_{\text{to}}) \in \boldsymbol{J}} \left(w_{j_{\text{from}}}^{a_k} w_{o_s^{j_{\text{from}}}}^{a_k} + w_{j_{\text{to}}}^{a_k} w'^{a_k}_{o_s^{j_{\text{to}}}} \right)$$

where o_s^i is the option of i and is included in a draft agreement s. The function is derived from a linear additive utility function, utilized in automated bilateral negotiation (Baarslag et al., 2016) and in Lewis et al. (2017).

3.3 Data Collection

We hired workers through Amazon Mechanical Turk to collect human-human dialogues. Only those based in the USA with at least 1,000 previous HITs and an approval rating of over 95% could join our experiments. Before each session, the workers read the task description and instructions for negotiating with the opponent[1]. During a negotiation, each worker could propose a draft agreement up to three times and was asked to send six messages or more in total to submit the proposal. We paid \$0.20 per dialogue and gave a $\$(\text{score} - 5)/5$ bonus if the score was more than $5/10$ to promote efficient negotiations.

3.4 Quantitative Comparison

Table 3 shows the quantitative comparison of three negotiation dialogue corpora. The vocabulary size is the largest in the CB dataset because it handles several categories of listed products. The JI and DN datasets focus on a single domain, and of the two, the former has the larger vocabulary size. The average number of turns per dialogue in the JI dataset is the largest of the three, though it has the smallest average number of words per turn. These statistics indicate that participants in the JI dataset likely had enough conversations to reach agreement.

Agreement Ratio The JI dataset had the highest agreement ratio of 92.9%, a sharp contrast with the values of 76.2% and 74.9% for the DN and CB datasets. This difference may be because the participants in the JI dataset could propose intermediate offers up to three times each, while those in the existing corpora could only submit one proposal per session.

Complexity of Negotiation Scenarios The JI dataset has far fewer Pareto optimal[3] solutions for agreements than the DN dataset, which can be ascribed to the following reasons: (1) the larger number of issues and options in the JI dataset, with 9920 possible solutions per dialogue, and (2) the introduction of an interdependent relationship that prevented the scoring function from following a standard linear additive utility function. As a result, participants in the JI dataset struggled to find better solutions and might have compromised with each other more often than in the DN dataset.

4 Task Description

Task We formally define the task of breakdown detection in negotiation dialogues. Let D be a negotiation dialogue between two negotiators, composed of $n \in \mathbb{N}$ turn's utterances $\{s_1, s_2, \ldots, s_n\}$, where each utterance s is a message from one of the negotiators and includes one or more sentences. Given D, the task is to label D as either a success (reaching an agreement: 0) or a breakdown (failing to find an agreement: 1).

Evaluation Metrics To evaluate the effectiveness of the different approaches, we employ area under curve (ROC-AUC) and confusion matrix (CM), both of which are based on Yamaguchi and

[3]When an agent's score cannot be improved without lowering the opponent's score, a solution is called Pareto optimal.

Fujita (2020). We also use average precision (AP) to consider the imbalanced nature of breakdown labels in negotiation datasets.

5 Methodology

This section introduces our breakdown detection approach using a dialogue act-based feature and ML models, including linear and NN-based models. The intuition that guides this feature is that because a breakdown dialogue should have distinct flow (e.g., many disagreements), focusing on the dialogue flow can help detect this type of breakdown.

5.1 Dialogue Act Extraction

Our dialogue acts and their extraction are based on He et al. (2018), but we made some changes in the extraction process to capture dialogue flow effectively. The process consists of two stages: (1) pattern matching and (2) filtering and alignment. The first step is almost identical to He et al. (2018), but the second is newly designed for this study.

Pattern Matching Given a dialogue turn, we extract dialogue acts according to the matching patterns (Table 4) using regular expressions[4]. If there is no matched pattern in it, an unknown tag <unk> is given. Note that because negotiators in the JI dataset can propose intermediate offers up to three times and because such offers are part of negotiations, we add corresponding dialogue acts whenever these offers are detected during conversations.

Filtering and Alignment He et al. (2018) only extracted one dialogue act per turn. However, because negotiators could send one or more sentences for each turn in the DN, CB, and JI datasets, there may have been two or more dialogue acts. To capture the dialogue flow in detail while matching noise due to the rule-based extraction is reduced, we filter extracted dialogue acts in a way that matches Figure 1, which only allows dialogue acts to appear in the designated order. If an illegal dialogue act follows a matched one, all remaining unmatched ones will be discarded. The constrained flow is motivated by an alternating-offer protocol (Rubinstein, 1982) utilized in automated negotiation (Baarslag et al., 2013b), where one agent proposes a draft agreement, and the other accepts it or makes a counter-offer. Although negotiation dialogues do not have a well-defined negotiation protocol, unlike the case of automated negotiation, we assume that human negotiators should follow an unwritten code to reach agreement with their opponents. Table 2 shows an example of extracted dialogue acts along with the text.

[4]If there are two or more matched patterns for the same dialogue act, only the first matched pattern is kept.

Dialogue act	Matching Pattern
<greet>	*hi, hello, yo, hey, hiya, howdy, how are you, good day, good afternoon, good morning*
<disagree>	Generic – *isn't, worse, bad, sorry, no, not, nothing, don't, can't, cannot, afraid, a lot lower/higher, too much/high/low* JI – An intermediate offer is rejected.
<agree>	*ok, okay, no problem, yes, great, perfect, thanks, gracias, thx, thank you, pleasure, fine, deal, cool, that works, that will work, that works, it will work, sounds good, very good, looks good, i can do*
<inquire>	*what, where, when, which, how's, how about, how does, do you, did you, will you, would you, could you, are you, do we, did we, could we, do i, let me know, ?*
<propose>	Generic – Any digits, *come down, highest, lowest, go higher/lower, i would like* DN – *ball(s), hat(s), book(s)* JI – A new intermediate offer is proposed.
<inform>	A previous utterance ends with <inquire> and its reply does not contain any other tags.

Table 4: Matching patterns for dialogue acts.

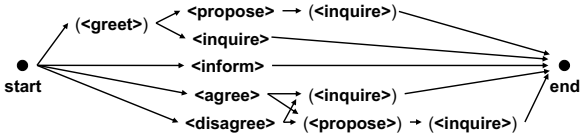

Figure 1: Constrained flow of each dialogue turn. Dialogue acts in parenthesis denote that they do not always appear in each turn.

5.2 Using Dialogue Act-based Features as Inputs for ML Models

Once we extract all features from a dialogue, we concatenate each turn with the addition of a separator tag <sep> to the head of each turn and an end tag <end> to the end of the dialogue. We then create an input vector for linear or NN-based models and use it to train the model. The input vector is produced as follows:

Linear Models We create a count vector by counting the number of each dialogue act per dialogue, including <unk>, <sep> and <end> tags.

NN-based Models We convert each extracted dialogue act into a one-hot representation $\boldsymbol{e} \in \mathbb{R}^{1\times 10}$, which includes a padding tag <pad>. We then concatenate all one-hot representations in time series per dialogue, which generates an input matrix $\boldsymbol{E} \in \mathbb{R}^{n\times 10}$, where n is the number of extracted dialogue acts, including padding.

6 Experimental Settings

6.1 Classification Models

We experiment with linear and NN-based models trained with either text-based or dialogue act-based features:

LR-BOW A logistic regression model trained with bag-of-words features weighted by TF-IDF.

GRU A GRU-based model with a linear layer on top of recurrent units. For text-based inputs, we used frozen pre-trained 300-dimensional word embeddding (GloVe) (Pennington et al., 2014). We also considered the model with a self-attention mechanism (GRU-Att) (Zhou et al., 2016).

BERT A pre-trained bidirectional encoder representations from transformers (BERT)-based model (Devlin et al., 2019) for only text-based inputs. We fine-tuned uncased $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$ models with one linear layer on the top of the [CLS] representation for binary classification.

Random A naive classifier that predicts negotiation outcomes by respecting training set's class distribution.

6.2 Data and Preprocessing

We employed three negotiation datasets compared in Table 3 for our experiments. The breakdown label of each dataset was assigned as follows. **DN**: A log has either a <disagree> or <no_agreement> tag inside an <output> tag. **CB**: A log does not have an offer price. **JI**: A "status" in a log is not "completed." For the CB and JI datasets, we removed short dialogues with less than three turns, as these are often labeled as breakdown and rarely include bargaining components, such as proposals. After the removal, the breakdown ratios of the CB and JI datasets were 18.9% and 4.9%. We preprocessed texts with lower-casing and inserted the <sep> and <end> tags into each dialogue, as in the dialogue act-based case. We tokenized the texts using spaCy[5]. For BERT, we used a pre-trained BERT tokenizer provided by the Transformers library (Wolf et al., 2020).

[5] https://spacy.io/

6.3 Implementation Details

We trained and tested models using stratified five-fold cross-validation. The model-specific implementation details are as follows:

Linear Model We implemented an LR-BOW model using Scikit-learn (Pedregosa et al., 2011) and trained it on Intel Core i5 (2.9 GHz - 6267U). We tested the n-gram combination of $\{(1,1),(1,2),(1,3)\}$. We applied L2 regularization and weight adjustments to make the weights inversely proportional to the labels in training data.

NN-based Models We set the maximum number of epochs to 100 for GRU-based models and 20 for BERT-based models, with early stopping. We further split the training folds into training (80%) and validation subsets (20%). We used the binary cross-entropy loss and optimized the models with an Adam optimizer (Kingma and Ba, 2014). We implemented the models using PyTorch (Paszke et al., 2019) and tuned their hyperparameters based on validation F_1[6]. For BERT-based models, we utilized the implementation provided by HuggingFace (Wolf et al., 2020). We trained and tested our models with NVIDIA Tesla V100 (SXM2 - 32GB).

7 Results and Analysis

7.1 Quantitative Results

Results in Existing Corpora We can observe from Table 5 that a fine-tuned $\text{BERT}_{\text{BASE}}$ model shows the best AP for the DN and CB datasets. Moreover, NN-based models with text-based features exhibit results that are comparable to those of the best-performing models in terms of AP, in the 95% confidence interval. The proposed approach (GRU_{TAG}) also showed comparable results for either AP or CM in both datasets. Although a logistic regression model with text-based features ($\text{LR-BOW}_{\text{TEXT}}$) produced poor results in terms of AP, it showed the best results for the pair of FN and TP and that of TN and FP in the DN and CB datasets, respectively.

Results in Proposed Corpus Our GRU-based models with dialogue act-based features (GRU_{TAG} and $\text{GRU-Att}_{\text{TAG}}$) showed by far the best AP of all

[6] Details concerning the hyperparameter selection are given in Appendix B.

Model	ROC-AUC	AP	TN	FP	FN	TP
DEALORNODEAL						
LR-BOW$_{TAG}$	.500 (n/a)	.238 (n/a)	1.00 (n/a)	.000 (n/a)	1.00 (n/a)	.000 (n/a)
GRU$_{TAG}$	.839 (.018)*	.766 (.031)*	.944 (.035)*	.056 (.035)*	.393 (.034)	.607 (.034)
GRU-Att$_{TAG}$	.834 (.012)*	.764 (.022)*	**.946** (.030)	**.054** (.030)	.406 (.014)	.594 (.014)
LR-BOW$_{TEXT}$	.838 (.024)	.745 (.031)	.891 (.011)	.109 (.011)	**.345** (.030)	**.655** (.030)
GRU$_{TEXT}$	.838 (.022)*	.772 (.031)*	.942 (.022)*	.058 (.022)*	.371 (.012)*	.629 (.012)*
GRU-Att$_{TEXT}$	.845 (.023)*	**.779** (.026)	.942 (.016)*	.058 (.016)*	.361 (.021)*	.639 (.021)*
BERT$_{BASE}$	.850 (.017)*	**.779** (.030)	.942 (.013)*	.058 (.013)*	.349 (.037)*	.651 (.037)*
BERT$_{LARGE}$	**.851** (.018)	.769 (.036)*	.940 (.011)*	.060 (.011)*	.354 (.036)*	.646 (.036)*
Random	.502 (.006)	.238 (.002)	.754 (.014)	.246 (.014)	.750 (.008)	.250 (.008)
CRAIGSLISTBARGAIN						
Model	ROC-AUC	AP	TN	FP	FN	TP
LR-BOW$_{TAG}$	.500 (n/a)	.189 (n/a)	1.00 (n/a)	.000 (n/a)	1.00 (n/a)	.000 (n/a)
GRU$_{TAG}$	.897 (.013)	.702 (.035)	.906 (.021)*	.094 (.021)*	.306 (.032)*	.694 (.032)*
GRU-Att$_{TAG}$	.893 (.016)	.679 (.035)	.894 (.037)	.106 (.037)	.312 (.050)	.688 (.050)
LR-BOW$_{TEXT}$	.874 (.013)	.685 (.024)	**.925** (.021)	**.075** (.021)	.398 (.029)	.602 (.029)
GRU$_{TEXT}$	.919 (.011)*	.755 (.033)*	.921 (.015)*	.079 (.015)*	.267 (.040)*	.733 (.040)*
GRU-Att$_{TEXT}$	**.920** (.014)	.737 (.025)*	.918 (.013)*	.082 (.013)*	**.261** (.040)	**.739** (.040)
BERT$_{BASE}$	**.920** (.008)	**.756** (.021)	.914 (.017)*	.086 (.017)*	.301 (.052)*	.699 (.052)*
BERT$_{LARGE}$	.910 (.017)*	.744 (.040)*	.919 (.003)*	.081 (.003)*	.299 (.033)*	.701 (.033)*
Random	.501 (.015)	.190 (.004)	.814 (.016)	.186 (.016)	.813 (.038)	.187 (.038)
JOBINTERVIEW						
Model	ROC-AUC	AP	TN	FP	FN	TP
LR-BOW$_{TAG}$	.500 (n/a)	.049 (n/a)	1.00 (n/a)	.000 (n/a)	1.00 (n/a)	.000 (n/a)
GRU$_{TAG}$	.902 (.016)*	**.418** (.035)	**.971** (.012)	**.029** (.012)	.646 (.102)*	.354 (.102)*
GRU-Att$_{TAG}$	**.915** (.014)	.416 (.076)*	.953 (.034)	.047 (.034)	**.582** (.186)	**.418** (.186)
LR-BOW$_{TEXT}$	.736 (.058)	.178 (.045)	.913 (.051)	.087 (.051)	.701 (.082)*	.299 (.082)*
GRU$_{TEXT}$	.539 (.083)	.093 (.024)	.966 (.032)*	.034 (.032)*	.937 (.031)	.063 (.031)
GRU-Att$_{TEXT}$	.547 (.089)	.086 (.017)	.964 (.027)*	.036 (.027)*	.922 (.065)	.078 (.065)
BERT$_{BASE}$	.705 (.059)	.172 (.072)	.951 (.040)	.049 (.040)	.802 (.111)*	.198 (.111)*
BERT$_{LARGE}$	.725 (.059)	.171 (.043)	.959 (.024)*	.041 (.024)*	.810 (.094)*	.190 (.094)*
Random	.515 (.025)	.053 (.005)	.951 (.006)	.049 (.006)	.921 (.055)	.079 (.055)

Table 5: Performance comparison for three negotiation dialogue datasets: Best mean results are in **bold**. Values in parenthesis represent standard deviations over the five test folds. Values marked with * are within the 95% confidence interval of the best score for a given metric. Confusion matrices are normalized on a set each of true negative (TN) and false positive (FP), and true negative (TP) and false negative (FN).

models and better results in other metrics. For text-based models, an LR-BOW$_{TEXT}$ model showed better results in terms of AP, FN, and TP than NN-based models. While text-based GRU models could not detect signs of breakdowns at all, BERT-based models could detect them with a TP ratio of 19.8% (base) and 19.0% (large).

Discussion First, dialogue act-based features only worked with sequential models. This result is in line with our key concept of capturing negotiation flow. Because the LR-BOW$_{TAG}$ model could not consider sequential information, it could not detect breakdowns at all. Second, an LR-BOW$_{TEXT}$ model worked well in all datasets, indicating that text-based features themselves contain breakdown information. However, this approach produced more misclassification for successful dialogues in the DN and JI datasets than other models, but it

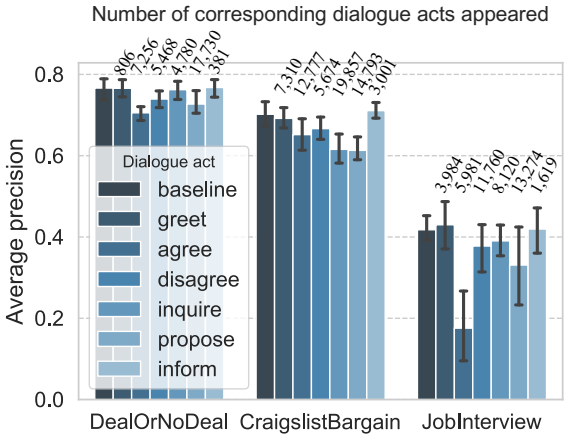

Figure 2: Classification performance comparison on five test folds when replacing a specific dialogue act with an unknown <unk> tag. Error bars denote the 95% confidence interval.

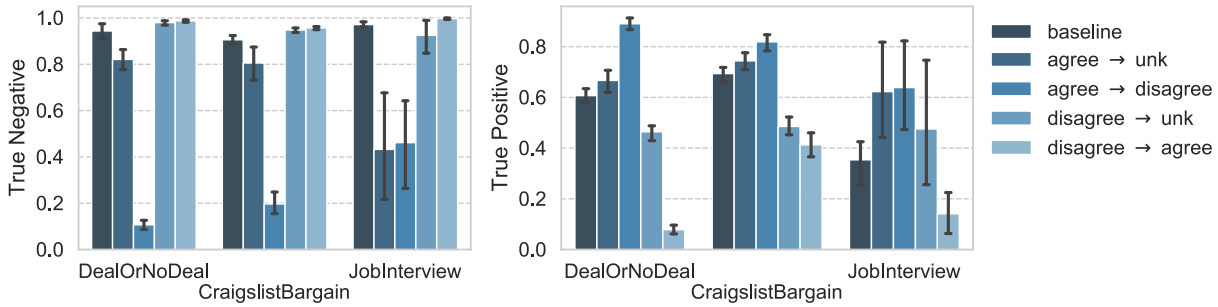

Figure 3: Performance comparison on five test folds when replacing <agree> and <disagree> tags with their counterpart or an <unk> tag. Error bars denote the 95% confidence interval.

could detect fewer breakdowns in the CB dataset. Because we intend to support human-human negotiation, accurate classification for both cases is vital to providing beneficial feedback to negotiators. Thus, the use of this approach is not helpful to our task. Third, NN-based models with text-based features did not perform well in the JI dataset. This was likely due to the far smaller breakdown ratio of 4.9% in the dataset compared to 23.8% and 18.9% in the DN and CB datasets. However, BERT-based models showed far better results than GRU-based ones in terms of the TP ratio. We hypothesize that BERT's rich contextualized information helped detect signs of breakdown.

7.2 Ablation Study

We conducted two ablation studies to better understand dialogue act-based input features. We first analyzed the importance of each dialogue act by replacing it with an unknown tag and tested with our best-performing model (GRU_{TAG}) over the five test folds. The <agree> tag was important for breakdown detection across the three corpora, despite its infrequency, especially in the DN and JI datasets (Figure 2). The frequent tag <propose> also played an important role in classification. By contrast, the <disagree> and <inquire> tags were not important except for the <inquire> tag in the CB dataset, possibly due to its highest frequency. Finally, the <greet> and <inform> tags were the least important in all datasets as these appeared less frequently and are not as closely related to breakdown as the others.

Next, we verified whether the GRU_{TAG} model captured the roles of <agree> and <disagree> tags in the breakdown detection task by replacing these tags with their counterpart or an <unk> tag (Figure 3). By replacing an <agree> tag with a <disagree> tag, we saw a rise in a TP ratio and a significant drop in a TN ratio compared to the baseline. When the <disagree> tag was replaced with an <agree> tag, the TN ratio slightly increased, while the TP ratio significantly decreased. These results suggest that the model properly took into account the roles of "<agree>" and "<disagree>" to some extent, and the number of such tags appeared played an important role in detecting a breakdown. While replacement with an <unk> tag also showed a similar trend, except with the <disagree> tag in the JI dataset, this was probably due to the relative increase of the counterpart.

FP (DN)	<sep> *i'd love to take a book and two hats off your hands* <sep> *hm, not many points for me but i'll agree to that.* <end>
	<sep> <propose> <sep> <disagree> <end>
FN (CB)	<sep> *hello, i am very interested in your car. however $12000 is out of my price range for a car that is 7 years old. i offer $6000 and i will pick up the car myself.* <sep> *there is no possible way i could go that low. i would take $11, 000* <sep> *that's fine, i will go elsewhere with my money.* <sep> *okay* <end>
	<sep> <greet> <propose> <sep> <disagree> <propose> <sep> <agree> <sep> <agree> <end>

Table 6: Examples of misclassified dialogues with extracted dialogue acts.

7.3 Error Analysis

Last, we conducted error analyses to examine the behavior of a GRU_{TAG} model and reveal its potential limitations. The first example is an FP sample from the DN dataset, where the model possibly focused on a <disagree> tag corresponding to *not*. The second one is an FN sample from the CB

dataset, in which the model might have focused on repetitive <agree> tags. We consider that the proposed approach could not cope with euphemistic phrases because of the rule-based dialogue act extraction. Thus, annotating negotiation corpora with dialogue acts will be an important research direction for more precise detection.

8 Conclusions and Future Work

This study proposed a job interview negotiation dialogue dataset with 2639 dialogues and increased complexities compared to existing datasets to help propel development of the study of human-human negotiation support and goal-oriented dialogue systems. We also proposed a dialogue act-based breakdown detection model that can focus on negotiation flow. Our approach (GRU_{TAG}) showed comparable results when used with existing datasets and better results for the proposed dataset than models trained with text-based features. In the future, we intend to explore another application of dialogue act-based features to related tasks, such as preference estimations. We will also utilize the proposed corpus in related tasks in human-human negotiation support and goal-oriented dialogue systems.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631. Association for Computing Machinery.

Muddasser Alam, Alex Rogers, and Sarvapali Ramchurn. 2013. Interdependent multi-issue negotiation for energy exchange in remote communities. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, pages 25–31.

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*, pages 2721–2727. European Language Resources Association.

Tim Baarslag, Reyhan Aydoğan, Koen V Hindriks, Katsuhide Fujita, Takayuki Ito, and Catholijn M Jonker. 2015. The automated negotiating agents competition, 2010–2015. *AI Magazine*, 36(4):115–118.

Tim Baarslag, Katsuhide Fujita, Enrico H. Gerding, Koen Hindriks, Takayuki Ito, Nicholas R. Jennings, Catholijn Jonker, Sarit Kraus, Raz Lin, Valentin Robu, and Colin R. Williams. 2013a. Evaluating practical negotiating agents: Results and analysis of the 2011 international competition. *Artificial Intelligence*, 198:73 – 103.

Tim Baarslag, Mark J. Hendrikx, Koen V. Hindriks, and Catholijn M. Jonker. 2016. Learning about the opponent in automated bilateral negotiation: A comprehensive survey of opponent modeling techniques. *Autonomous Agents and Multi-Agent Systems*, 30(5):849–898.

Tim Baarslag, Koen Hindriks, and Catholijn Jonker. 2013b. Acceptance conditions in automated negotiation. In *Complex Automated Negotiations: Theories, Models, and Software Competitions*, pages 95–111. Springer.

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. Association for Computational Linguistics.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

Heriberto Cuayáhuitl, Simon Keizer, and Oliver Lemon. 2015. Strategic dialogue management via deep reinforcement learning. *CoRR*, abs/1511.08099.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. 2016. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 3146–3150. European Language Resources Association.

Chiori Hori, Julien Perez, Ryuichiro Higashinaka, Takaaki Hori, Y-Lan Boureau, Michimasa Inaba,

Yuiko Tsunomori, Tetsuro Takahashi, Koichiro Yoshino, and Seokhwan Kim. 2019. Overview of the sixth dialog system technology challenge: Dstc6. *Computer Speech & Language*, 55:1 – 25.

Takayuki Ito, Mark Klein, and Hiromitsu Hattori. 2006. A negotiation protocol for agents with nonlinear utility functions. In *Proceedings of the Twenty-First AAAI Conference on Artificial Intelligence*.

Kosui Iwasa and Katsuhide Fujita. 2018. Prediction of nash bargaining solution in negotiation dialogue. In *PRICAI 2018: Trends in Artificial Intelligence*, pages 786–796. Springer International Publishing.

Catholijn Jonker, Reyhan Aydogan, Tim Baarslag, Katsuhide Fujita, Takayuki Ito, and Koen Hindriks. 2017. Automated negotiating agents competition (anac). In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 5070–5072.

A. Kardan and H. Janzadeh. 2008. A multi-issue negotiation mechanism with interdependent negotiation issues. In *Proceedings of the Second International Conference on the Digital Society*, pages 55–59.

Ralph L. Keeney and Howard Raiffa. 1993. *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*. Cambridge University Press.

Simon Keizer, Markus Guhe, Heriberto Cuayáhuitl, Ioannis Efstathiou, Klaus-Peter Engelbrecht, Mihai Dobre, Alex Lascarides, and Oliver Lemon. 2017. Evaluating persuasion strategies and deep reinforcement learning methods for negotiation dialogue agents. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 480–484. Association for Computational Linguistics.

Wolfgang Ketter, John Collins, and Prashant Reddy. 2013. Power tac: A competitive economic simulation of the smart grid. *Energy Economics*, 39:262 – 270.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Vasily Konovalov, Ron Artstein, Oren Melamud, and Ido Dagan. 2016. The negochat corpus of human-agent negotiation dialogues. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 3141–3145. European Language Resources Association.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453. Association for Computational Linguistics.

Raz Lin and Sarit Kraus. 2010. Can automated agents proficiently negotiate with humans? *Commun. ACM*, 53(1):78–88.

Johnathan Mell, Jonathan Gratch, Tim Baarslag, Reyhan Aydoğan, and Catholijn M. Jonker. 2018. Results of the first annual human-agent league of the automated negotiating agents competition. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 23–28. Association for Computing Machinery.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543. Association for Computational Linguistics.

Volha Petukhova, Harry Bunt, and Andrei Malchanau. 2017. Computing negotiation update semantics in multi-issue bargaining dialogues. In *Proceedings of SEMDIAL 2017 (SaarDial) Workshop on the Semantics and Pragmatics of Dialogue*, pages 87–97.

Volha Petukhova, Christopher Stevens, Harmen de Weerd, Niels Taatgen, Fokie Cnossen, and Andrei Malchanau. 2016. Modelling multi-issue bargaining dialogues: Data collection, annotation design and corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 3133–3140. European Language Resources Association.

Howard Raiffa, John Richardson, David Metcalfe, et al. 2002. *Negotiation analysis: The science and art of collaborative decision making*. Harvard University Press.

Ariel Rubinstein. 1982. Perfect equilibrium in a bargaining model. *Econometrica*, 50:97–109.

Leigh L. Thompson, Jiunwen Wang, and Brian C. Gunia. 2010. Negotiation. *Annual Review of Psychology*, 61(1):491–515.

David Traum, Stacy C. Marsella, Jonathan Gratch, Jina Lee, and Arno Hartholt. 2008. Multi-party, multi-issue, multi-strategy negotiation for multi-modal vir-

tual agents. In *Proceedings of the Eighth International Conference on Intelligent Virtual Agents*, pages 117–130. Springer.

Minhong Wang, Huaiqing Wang, Doug Vogel, Kuldeep Kumar, and Dickson K. W. Chiu. 2009. Agent-based negotiation and decision making for dynamic supply chain formation. *Engineering Applications of Artificial Intelligence*, 22(7):1046–1055.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Atsuki Yamaguchi and Katsuhide Fujita. 2020. Breakdown detection in negotiation dialogues (student abstract). In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 13969–13970.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212. Association for Computational Linguistics.

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. Association for Computational Linguistics.

A Job Interview Negotiation Dataset

Here, we introduce the negotiation interface and negotiation procedures. Our dataset and negotiation interface are available at https://github.com/gucci-j/negotiation-breakdown-detection.

A.1 Negotiation Interface

We developed an online negotiation interface for our job-interview negotiation, which implemented all mathematical settings such as preferences and a scoring function discussed in the body of the paper. Figure 4 shows the screenshot of our negotiation interface.

At the beginning of each negotiation session, the interface generates negotiators' preferences and displays them next to the corresponding issues and options so that the negotiators can easily understand which issue and option is important for them.

During the session, whenever the negotiators select a new solution, the interface calculates the score of its solution according to the scoring function described in Subsection 3.2 and displays it with the corresponding evaluation. The evaluation is based on Table 7 and intended for providing feedback to the negotiators to promote a better agreement.

At the end of the session, the interface stores the log that consists of the preferences of the participants, dialogue history, proposed offers and settled agreement in `json` format.

Score	Evaluation
< 50	Very bad
< 60	Bad
< 70	Fair
< 80	Good
< 90	Very good
≥ 90	Excellent

Table 7: Correspondence table between the score and the evaluation.

A.2 Negotiation Procedures

Before entering a negotiation session, each negotiator reads the instruction page that describes the outline of the negotiation, its procedures and some precautions (e.g., the maximum number of proposals per negotiator).

During the session, the negotiators can talk to their opponent using the left-hand side of the negotiation interface (Figure 4), while they can select an option for each issue in the right-hand side of the interface. Besides, they can also check the current score, its evaluation and estimated HIT reward for the selected options.

When the negotiators believe that they had sufficient discussion, they can propose a draft agreement by clicking the "PROPOSE" button shown in the bottom-left side of the interface. Once it is sent to the opponent, the opponent can check its details and score with the "ACCEPT" button shown on the interface. If the opponent clicks the button, the negotiation is regarded as successful. Otherwise, the negotiation continues until both the sides exceed the maximum number of propositions. If exceeding the limit, the negotiation is regarded as a breakdown, and the score of each negotiator is recorded as zero.

Hyperparameter	Value or search space
Maximum training epochs	100
Mini-batch size	64
Adam β_1	0.9
Adam β_2	0.999
Learning rate	$[10^{-5}, 10^{-2}]$
Early stopping patience value	8
Number of GRU layers	$[1, 4]$
Number of GRU hidden units	$[64, 256]$
Bidirectional	`True` or `False`
Recurrent dropout rate	$(0.0, 1.0)$
Classifier dropout rate	$(0.0, 1.0)$

Table 8: Hyperparameters and search space for GRU-based models. If "bidirectional" is `True`, a model becomes a bidirectional GRU.

B Hyperparameter Tuning

Linear Models For the DN dataset, n-gram combination of $(1, 3)$ (uni-gram, bi-gram, and tri-gram) was chosen. For the CB dataset, that of $(1, 2)$ (uni-gram and bi-gram) was selected. For the JI dataset, that of $(1, 1)$ (uni-gram) was chosen. Since none of the models trained with dialogue act-based features did not work, these have no optimal n-gram combinations.

Neural Network-based Models We tuned the hyperparameters of all NN-based models employed in our experiments using the Optuna framework (Akiba et al., 2019). We split training folds into training (80%) and validation (20%) subsets. We tested 100 hyperparameter combinations and evaluated their performance based on F_1 in each validation subset. Tables 8 and 9 show the hyperparameters and search space for GRU and BERT-based models, respectively.

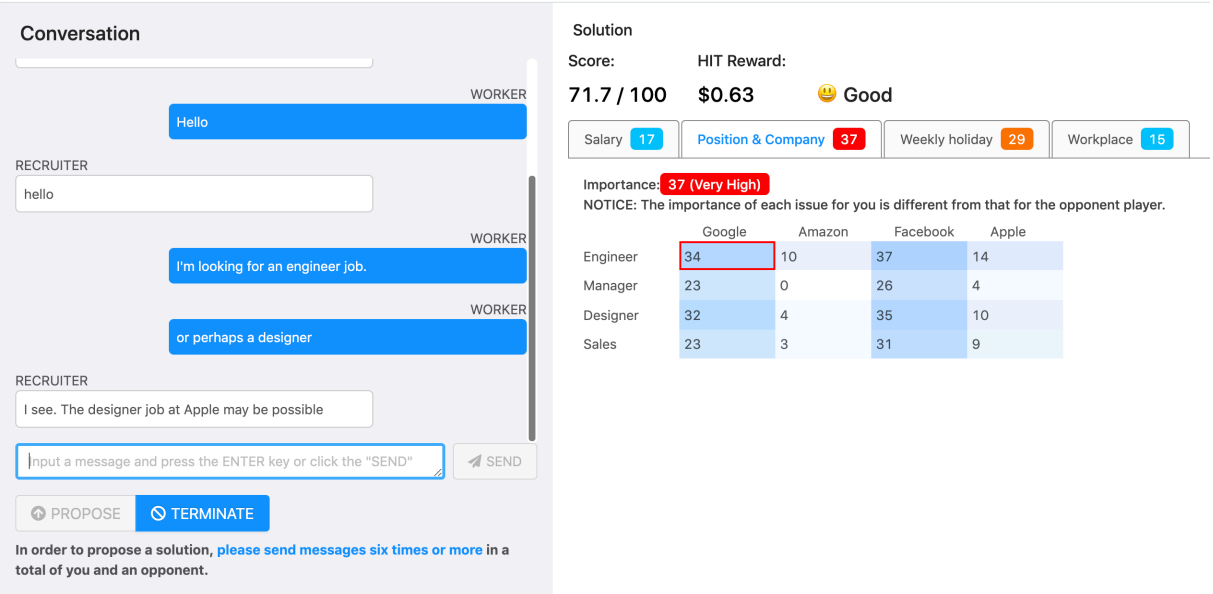

Figure 4: Negotiation interface used for the JI dataset. Each value shown next to an issue or an option denotes its importance for a negotiator. The score and importance of each issue and option were calculated by the interface based on the mathematical settings discussed in the body of the paper. Note that the score shown on the interface are multiplied by ten for the ease of players' understanding.

Hyperparameter	**Value or search space**
Maximum training epochs	20
Mini-batch size	16 ($\text{BERT}_{\text{LARGE}}$) 32 ($\text{BERT}_{\text{BASE}}$)
Adam β_1	0.9
Adam β_2	0.999
Maximum sequence length	196 (CB and JI datasets) 128 (DN dataset)
Learning rate for pre-trained layers	$[10^{-6}, 10^{-3}]$
Learning rate for an additional dense layer	$[10^{-5}, 10^{-2}]$
Learning rate scheduler	{"get cosine schedule with warmup," "get constant schedule with warmup," "get linear schedule with warmup"}
Warmup steps	$[1, 120]$
Early stopping patience value	3
Dropout rate	$(0.0, 1.0)$
Gradient accumulation steps	10 ($\text{BERT}_{\text{LARGE}}$) 5 ($\text{BERT}_{\text{BASE}}$)

Table 9: Hyperparameters and search space for BERT-based models. Each scheduler name corresponds to the one in the Transformers library (Wolf et al., 2020) by replacing blanks with "_".