Negotiation Dialogue System Using a Deep Learning-Based Parser

Kenjiro Morimoto, Katsuhide Fujita and Ken Watanabe
Tokyo University of Agriculture and Technology, Tokyo, Japan

Keywords: Negotiation Dialogue System, Deep Learning, Parser.

Abstract: In recent years, there has been substantial research on negotiation dialogue agents. A notable study introduced a method that decoupled strategy from generation using dialogue acts that encapsulated the intent behind utterances. This approach has enhanced both the task success rate and the human-like quality of the generated responses. However, the rule-based implementation of the parser limits the types of sentences it can process for dialogue acts. Thus, this paper presents annotated training data based on the proposed dialogue acts and introduces a deep learning-based parser. The deep learning-based parser achieved a dialogue act classification accuracy of approximately 83% and effectively reduced the occurrence of unknown dialogue acts. Additionally, negotiation dialogue systems using deep learning-based parsers have demonstrated improved performance in terms of utility and fairness.

1 INTRODUCTION

Negotiation is a crucial skill in human communication for resolving conflicts and achieving beneficial agreements (Fisher et al ., 2011) (Lewicki et al ., 2011). Recently, there has been extensive research on negotiation dialogue systems, which focus on creating intelligent dialogue agents capable of negotiating with humans through natural language (Zhan et al ., 2020) (Basave et al ., 2016). The goal of these negotiation dialogue systems is to facilitate conflict resolution and enhance mutual benefits by generating context-appropriate negotiation dialogues.

One study on negotiation dialogue systems presents a structural model that incorporates dialogue actions, known as dialogue acts, into the natural language understanding and generation process (He et al ., 2018). This study proposed a framework comprising three modules: a parser that transforms input utterances into dialogue acts, a manager that produces response dialogue acts based on dialogue act history and the dialogue scenario, and a generator that converts the dialogue acts generated by the manager into natural language responses. This approach differentiates the formulation of negotiation strategies from the dialogue generation process, enhancing human-like interaction and increasing task completion rates. However, this framework has room for improvement because its parser is rule-based and unable to assign dialogue acts accurately based on the meaning of natural language sentences.

The aim of this study is to introduce a more accurate method for estimating dialogue acts using deep learning for a parser that assesses dialogue acts corresponding to input sentences. While rule-based methods offer the benefits of high explainability and ease of implementation, they struggle with data that fall outside predefined rules, and it is challenging to establish rules for all possible scenarios. In this paper, we present innovative dialogue acts and a deep learning-based method for a parser. We further illustrate the effectiveness of the proposed method through comparative experiments with previous studies. The parser model is created by fine-tuning several pretrained models using training data annotated with the proposed dialogue acts. We integrate the proposed parser along with the previous method into negotiation dialogue systems and conduct human-agent negotiation experiments with participants to assess the deep learning-based parser. Our findings indicate that a deep learning-based parser can learn data features that a rule-based parser cannot handle and can accurately infer dialogue acts across a broader range of data. Furthermore, a negotiation dialogue system that incorporates a deep learning-based parser has been shown to enhance performance in areas such as utility and fairness indices.

Morimoto, K., Fujita, K. and Watanabe, K.
Negotiation Dialogue System Using a Deep Learning-Based Parser.
DOI: 10.5220/0013245200003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 1*, pages 135-143
ISBN: 978-989-758-737-5; ISSN: 2184-433X

2 RELATED WORKS

2.1 Dialogue Act

A dialogue act is a structural model that represents the actions in a dialogue. Its purpose is to classify each utterance based on the speaker's intention (Żelasko et al ., 2021). Each dialogue act comprises an intent that conveys the meaning of the utterance and an argument (e.g., price). Dialogue acts are essential for understanding spoken language, particularly in the fields of linguistics and artificial intelligence. In artificial intelligence, it is crucial to establish appropriate dialogue acts and perform annotation according to the type of dialogue data being processed. Below are examples of dialogue acts found in the negotiation dialogues examined in this study.

- Utterance = Hi, I'm interested in your bike.
- Dialogue Act = greet
- Utterance = I have it listed for $220.
- Dialogue Act = init-price (220).

2.2 Craigslist Negotiation Dataset

CRAIGSLISTBARGAIN is a dataset comprising price negotiation conversations for items listed on Craigslist, an American classified advertising community site (He et al ., 2018). In contrast to many earlier negotiation dialogue datasets that were gathered from limited dialogue domains, such as games (Lewis et al ., 2017) (Asher et al ., 2016), CRAIGSLISTBARGAIN includes negotiation dialogues that feature side offers and casual discussions, offering scenarios that closely resemble real-life negotiation settings.

In a two-party negotiation in CRAIGSLISTBARGAIN, two agents take on the roles of buyer and seller, respectively. Each agent receives photos, descriptions, and listed prices of items available on Craigslist. The seller aims to negotiate a sale at the listed price, while the buyer seeks to purchase at an undisclosed target price. Either agent has the discretion to make a price offer, which can be accepted or rejected by the other party. Additionally, agents have the option to terminate negotiations and end the task without reaching an agreement. The negotiation scenarios are centered around the six most popular categories of Craigslist posts: housing, furniture, cars and bikes, phones, and electronics. This dataset consists of 6,682 person-to-person conversations collected via Amazon Mechanical Turk. We use the CRAIGSLISTBARGAIN dataset for the learning and evaluation experiments of the negotiation dialogue system in this study.

2.3 BERT

Bidirectional Encoder Representations from Transformers (BERT) is a natural language processing model introduced by Google in 2018 (Devlin et al ., 2018). Unlike traditional pretraining approaches that focus on unidirectional context, BERT uses two pretraining methods to analyze context in both directions. This pretraining process uses a substantial collection of unlabeled sentences, allowing for fine-tuning the model for various natural language processing tasks simply by adding an output layer. The paper reports enhanced accuracy across 11 different natural language processing tasks, and this technology continues to be used today in numerous applications, including search engines and chatbots.

2.4 Modular Framework

The objective of a negotiation dialogue system is to analyze a series of utterances $x_1, x_2, \ldots, x_{t-1}$ associated with a dialogue scenario c and produce a distribution for the response utterance x_t. In this research, we use a framework (He et al ., 2018) that incorporates dialogue acts into the strategies of traditional goal-oriented dialogue systems (Young et al ., 2013). Figure 1 shows this framework.

The framework shown in Figure 1 consists of the following three types of modules.

1. A parser that transforms input utterances into dialogue acts by leveraging the dialogue history $x_{<t}$, the dialogue act history $z_{<t}$, and a negotiation scenario c.
2. A manager that predicts a dialogue act z_t as a response to input, using the dialogue act history $z_{<t}$ and the negotiation scenario c.
3. A generator that takes the predicted dialogue act z_t from the manager and converts it, along with the dialogue history $z_{<t}$, into a natural language response x_t.

2.4.1 Parser

The Modular framework focuses on dialogue acts, which are composed of intents that convey the purpose of an utterance and arguments tailored to specific scenarios. For instance, the utterance "I have it listed for $220" is categorized as a dialogue act with the intent init-price and the argument price=200. Dialogue acts serve as structural models that offer a high-level understanding of a sentence rather than aiming to capture its entire meaning.

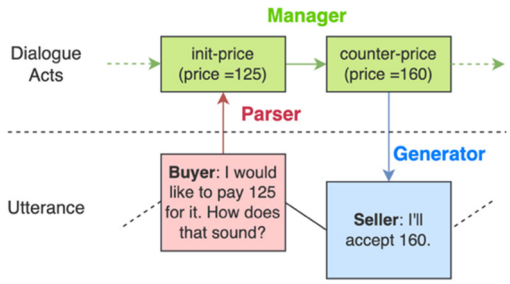

Figure 1: Modular framework (He et al ., 2018).

Consequently, this framework uses a rule-based parser that uses regular expressions and if–then rules. Rule-based systems process information based on artificially established rules, granting them the advantage of high explainability and straightforward implementation. The parser identifies keywords from the utterance and aligns them with predefined keyword patterns. Matching rules are organized as a sequential list. When multiple patterns match, the first identified intent is chosen. If no patterns apply, an unknown intent is returned. The intents used and their corresponding matching patterns from previous studies (He et al ., 2018) are presented in Table 1.

2.4.2 Manager

The manager's role is to identify the appropriate dialogue act for an utterance that the agent should choose at each time step t, considering the history of dialogue acts $z_{<t}$ and the dialogue scenario c. The manager is trained through supervised learning, enhanced by an attention mechanism, and also incorporates reinforcement learning with three distinct reward functions.

In supervised learning, we aim to maximize the likelihood of the training data based on the provided dialogue act history $z_{<t}$ and dialogue scenario c, while also learning the transition probability $p_\theta(z_t|z_{<t}, c)$. During the agent's listening turn, the Long Short Term Memory (LSTM) encodes the incoming dialogue acts. Conversely, during the agent's speaking turn, a different LSTM decodes the tokens in the dialogue act. The hidden layer's state is maintained across conversations to ensure a comprehensive dialogue history.

In reinforcement learning, three reward functions—Utility, Fairness, and Length—are optimized using the policy gradient method. Utility is the agent's self-interested objective, designed as a linear function of the final interaction price. Buyers achieve a utility of 1 at the target price, while sellers reach a utility of 1 at the list price. Additionally, both agents have zero utility at the midpoint between the list price and the buyers' target price. Fairness focuses on equalizing the utilities for both buyers and sellers. Length measures the number of utterances in a dialogue, promoting the agent to maintain the conversation for as long as possible. If no agreement is made, the reward assigned is −1.

Table 1: Intent used in the previous study (He et al ., 2018).

Intent	Matching Patterns
intro	Hi, hello, hey, hiya, howdy, how are you, interested
inquiry	starts with an interrogative word (e.g., what, when, where) or particle (e.g., do, are)
inform	previous dialogue act was inquire
init-price	first price mention
vague-price	No price mention and comedown, highest, lowest, go higher/lower, too high/low
counter-price	New price detected
insist	The same offer as the previous one is detected
disagree	No, not, n't, nothing, don't
agree	Not disagree and, ok, okay, great, perfect, deal, that works, I can do that
unknown	Does not match any rule

2.4.3 Generator

The generator's primary function is to transform dialogue act predictions (z_t) made by the manager into natural language utterances (x_t). Previous research has used a search-based approach to implement the generator. The search-based approach leverages a database of templates generated from the training dataset's utterances, which have been analyzed by a parser. Each utterance is converted into a template by replacing specific words with placeholders based on the corresponding dialogue act. For instance, the utterance "Would you take $705 for it?" is transformed into the template "Would you take [price] for it?" by substituting the numerical value "$705" with the placeholder [price].

Natural language utterances are generated by assessing the similarity between the template context and the current context. We represent each context as a BOW vector weighted by TF-IDF. Similarity is then calculated by taking the dot product of the two context vectors. To enhance the diversity of generated utterances, we select one utterance from the top K candidates guided by a distribution derived from a 3-gram language model trained on the training data.

3 PROPOSED DIALOGUE ACTS AND DEEP-LEARNING BASED PARSER

3.1 Proposed Dialogue Acts

Table 2 presents the results of classifying utterances from the CRAIGSLISTBARGAIN dataset using a rule-based parser from a previous study (He et al ., 2018). The table shows that the rule-based parser assigns "unknown" intent to approximately 25% of the utterances. This indicates that the parser is unable to classify these utterances. In the negotiation dialogue system framework, sentence understanding and generation rely on dialogue acts. The high "unknown" intent rate signifies that approximately 25% of the generated responses may not accurately reflect the intended dialogue act. This indicates a potential weakness in the system's ability to effectively understand and respond to user input.

To enhance the classification accuracy of utterances and assign them to appropriate intents, we introduce two new intents: "supplemental" and "thanks." The "supplemental" intent signifies the provision of additional information that contributes to the negotiation process. It may encompass detailed item descriptions, personal stories, or other relevant details. In negotiation dialogues, supplemental information often plays a crucial role in achieving a favorable outcome. For instance, highlighting an item's appealing features and associated benefits can increase the price, while disclosing one's financial situation and reasons for wanting to purchase can lead to a price reduction. While supplemental explanations in response to partner inquiries are categorized as "inform," spontaneous information sharing is often classified as "unknown." Introducing a supplemental dialogue act allows for utterances unrelated to the negotiation but supportive of it. For example, "Thanks" signifies gratitude toward the partner. Because negotiation dialogues involve human communication, they do not necessarily conclude immediately after agreement. In many instances, after expressing intent to reach an agreement, expressions of gratitude toward the partner are observed. Existing intent classifications lacked the ability to capture these expressions of gratitude, and most of them were classified as "unknown." By adding "thanks," we can address communication aspects beyond negotiations. Below are examples of utterances classified as "supplemental" and "thanks".

Table 2: Intent classification with a rule-based parser.

Intent	# of utterances	% of total # of utterances
unknown	9592	24.793
counter-price	7738	20.001
inquiry	5056	13.069
init-price	4629	11.965
intro	4611	11.918
inform	2321	5.999
disagree	2027	5.239
agree	1896	4.901
insist	432	1.117
vague-price	386	0.988
Total	38688	100

- Example of Supplemental
  Utterance = I can afford to pay $72 for it.
  Dialogue Act = init-price
  Utterance = This is antique, so although it is used, it is a very good bookcase.
  Dialogue Act = supplemental
- Example of Thanks
  Utterance = Ok, I can accept $12.
  Dialogue Act = agree
  Utterance = Great, thanks!
  Dialogue Act = thanks

In this study, we use a total of 12 intents, including those listed in Table 1, along with the additional intents "supplemental" and "thanks."

3.2 Annotation

Our study requires training data to develop a deep learning model capable of estimating intent. To prepare these data, seven individuals, encompassing both members of the general public and university students, annotated each of the 5,987 dialogues in the CRAIGSLISTBARGAIN dataset. To maintain consistent annotation quality across workers, we established classification criteria for each intent. Utterances in the dataset were then read according to these criteria to determine their respective intents. To enhance efficiency and minimize typographical errors, aliases were assigned to intents. Annotators used these aliases during the process, with subsequent conversion of all aliases to their original intents after annotation completion. The classification criteria for intents in annotation are shown in Table 3.

Table 3: Classification criteria for annotation.

Intent (Alias)	Classification criteria
intro (g)	Utterances that don't include a price indication and include words of greeting (e.g., hi, hello, hey, good day) or indicating a willingness to negotiate (e.g., interested in, do you have any question?, I want to buy)
Inquiry (q)	Utterances beginning with an interrogative word (e.g., what, when, where) or particle (e.g., do, are), or question to a partner
Inform (f)	Responses to questions
init-price (p)	Utterances that include a price indication and are the first price offer in the dialogue
vague-price (v)	Price negotiation utterances that ask for a price increase or reduction without a price listed
counter-price (c)	Price offer utterances that includes a price indication, follows "init-price", and don't fall under "insist", "agree", or "disagree"
insist (i)	Utterances proposing the same price as the previous one, or asserting the legitimacy of the price proposed
disagree (d)	Utterances to decline a price offer from a partner
agree (a)	Utterances to accept a price offer from a partner
supplemental (s)	Utterances that provide supplementary information to advance negotiations, such as detailed information of an item, and don't fall under "inform"
thanks (t)	Utterances expressing gratitude to a partner (e.g., thanks, thank you)
unknown (u)	Utterances that cannot be classified into any intent

Following the completion of annotation by each worker, the content undergoes a review and correction process. Each utterance is annotated by one worker and subsequently checked and corrected by a separate worker. Once all utterances have been annotated and reviewed, they are consolidated to form the new training data.

3.3 Learning Framework: Deep Learning-Based Parser

In this study, we propose a deep learning-based parser by fine-tuning BERT, a prominent Transformer encoder model (Vaswani et al ., 2017) that is well-suited for text classification tasks.

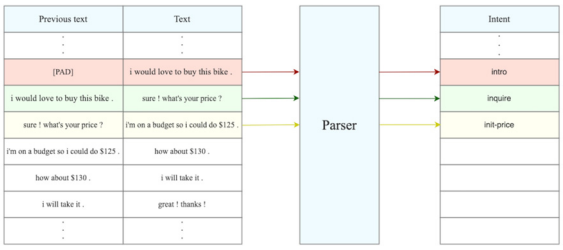

Figure 2: Inference process flow of a deep learning-based parser.

For the implementation using Hugging Face Transformers, Bert-base-uncased was chosen as the pretrained model, with the proposed dataset serving as the training data. The tokenizer associated with the selected model was used to tokenize the training data. Recognizing the interactive nature of dialogue, including negotiation dialogue where utterances frequently rely on preceding partner statements, the tokenizer input was structured as a two-sentence input. This approach considers not only the utterance to be classified but also the immediately preceding utterance. Utterance flow is segmented into individual dialogues. For the initial utterance in a dialogue, the [PAD] token is provided as the preceding utterance. Fine-tuning is then performed using the Trainer class. Figure 2 shows the inference processing flow used by the parser in this study. To facilitate performance comparison, parsers based on ALBERT (Lan et al ., 2020), DistilBERT (Sanh et al ., 2019), and RoBERTa (Liu et al ., 2019), all derivatives of BERT, were also developed. For these implementations, albert-base-v1, distilbert-base-uncased, and RoBERTa-base were selected as the respective pretrained models.

4 PERFORMANCES OF DEEP LERANING-BASED PARSERS

4.1 Inference Using Deep Learning-Based Parsers

We fine-tuned four pretrained models (BERT, ALBERT, DistilBERT, and RoBERTa) using annotated training data and evaluated the resulting models through cross-validation.

Tables 4 and 5 present the performance comparison results of fine-tuning for each pretrained model, highlighting the best results for each evaluation metric in bold. Table 4 shows that all pretrained models achieved a classification accuracy rate of approximately 83%. RoBERTa exhibited the

Table 4: Parser performance comparison.

Model	Accuracy (%)	Train runtimes (s)
BERT	83.456	1.451×10^4
ALBERT	83.196	0.898×10^4
DistilBERT	82.959	$\mathbf{0.810 \times 10^4}$
RoBERTa	**83.836**	1.563×10^4
Rule-base	43.960	-

Table 5: Intent classification evaluation by F1 score.

	BERT	ALBERT	DistilBERT	ROBERTa	# of Utterances
intro	0.886	0.885	0.881	**0.888**	1426
inquiry	**0.906**	0.905	0.902	**0.906**	1877
inform	**0.900**	0.895	0.896	**0.900**	1427
init-price	**0.833**	**0.833**	0.822	0.830	1139
vague-price	0.618	**0.627**	0.596	0.622	316
counter-price	0.863	0.849	0.853	**0.864**	1855
insist	0.174	0.115	0.131	**0.187**	111
disagree	0.535	0.505	0.502	**0.540**	152
agree	0.837	0.824	0.833	**0.840**	1147
supplemental	**0.596**	0.594	0.577	0.591	380
thanks	0.758	**0.775**	0.766	0.767	363
unknown	**0.302**	0.248	0.260	0.294	95
macro avg	0.684	0.671	0.668	**0.686**	10288

highest accuracy, surpassing DistilBERT, which had the lowest accuracy, by approximately 0.9%. Conversely, DistilBERT exhibited the shortest training runtimes, approximately half that of RoBERTa. ALBERT and DistilBERT, both lightweight BERT variations, maintained accuracy rates in 1% of the original BERT while considerably reducing training runtimes. The BERT derivative models had an accuracy variation of ±0.5%. However, the lightweight versions, such as DistilBERT, demonstrated significantly faster training times. Therefore, for larger datasets, using ALBERT or DistilBERT is recommended owing to their improved efficiency.

Table 5 shows that RoBERTa achieved the highest F1 score across all seven dialogue acts, outperforming all other models. Notably, BERT attained the highest F1 scores for supplemental and unknown intents, ALBERT for vague-price and thanks intents, and both BERT and ALBERT for init-price. This indicates that the classification of easier dialogue acts may be influenced by the specific pretrained model used. Conversely, DistilBERT consistently performed below all other models across all evaluation metrics.

Examining individual dialogue acts, those with over 1,000 data points consistently achieved F1 values exceeding 0.8 across all models. Conversely, the remaining six dialogue acts with fewer data points exhibited F1 values below 0.8, suggesting a decrease in classification accuracy for less frequent dialogue acts in negotiation dialogues. The F1 values for "insist" and "unknown" were particularly low, at 0.187 and 0.302, respectively. "Insist" represents a dialogue act aimed at reiterating a previous price offer. Given the input method used in this study, which involved inputting two sentences (the current utterance and the previous utterance), it is possible that this input method may have resulted in numerous misclassifications owing to the inability to fully capture contextual information. Because "unknown" encompasses a collection of sentences that resist classification, its features may not have been adequately learned during training, potentially contributing to the reduced classification accuracy. Misclassifications of sentences that should be classified as "unknown" into other dialogue acts may occur because the model learns features beyond human comprehension through deep learning, thereby accurately classifying them into the correct dialogue act. This presents a potential advantage of deep learning-based parsers, particularly relevant to negotiation dialogue systems. However, for the "insist" intent, it is necessary to either increase the number of training samples classified as "insist" or merge it with other dialogue acts to increase classification accuracy.

In conclusion, the pretrained models best suited for deep learning-based parsers are ALBERT, which demonstrates low computational complexity and execution time alongside high classification accuracy for the data-sparse dialogue acts of "vague-price" and "thanks," and RoBERTa, which exhibits the highest overall classification accuracy.

4.2 Results of Inference Using Deep Learning-Based Parsers

Table 5 presents the classification results of sentences in the CRAIGSLISTBARGAIN dataset using the deep learning-based parser. For this parser, a model based on RoBERTa, which achieved the highest accuracy rate, was used. Comparing Table 5 with Table 1 reveals that our proposed dialogue acts successfully reduced the proportion of sentences classified as "unknown" from 24.793% to 0.772%. Both "counter-price" and "inquiry" accounted for over 18% of the total utterance count, representing a substantial portion of all utterances. In the context of price negotiation dialogues, "counter-price" is deemed essential because price offers form the central theme of the conversation. However, because all questions

Table 6: Intent classification using a deep learning-based parser.

Intent	# of utterances	% of total # of utterances
unknown	9592	24.793
counter-price	7738	20.001
inquiry	5056	13.069
init-price	4629	11.965
intro	4611	11.918
inform	2321	5.999
disagree	2027	5.239
agree	1896	4.901
insist	432	1.117
vague-price	386	0.998
Total	38688	100

are currently classified under "inquiry," there is potential for further subdivision. "Disagree" and "insist" constituted a relatively small proportion of the total utterances, accounting for 1.068% and 0.677%, respectively. Owing to the tendency of deep learning-based parsers to exhibit lower classification accuracy for classes with limited training data, integrating "insist" and "disagree" into other dialogue acts would be advantageous. While integrating "disagree" poses a challenge owing to its frequent occurrence in negotiation scenarios, "insist" shares considerable similarities with "counter-price" and "vague-price," suggesting a potential solution of merging it into these two dialogue acts.

5 EVALUATIONS

5.1 Task

We evaluate our approach using the CRAIGSLISTBARGAIN task (Section 2.2), where a buyer and a seller negotiate the price of an item based on the information listed on Craigslist. In this task, the only argument in the dialogue act is the price.

5.2 Models

This study compares two models: a rule-based parser commonly used in previous research and our proposed deep learning-based parser. Both methods are applied to parse the dataset incorporating the newly annotated dialogue acts (Section 3.2).

The parser outputs are then used to generate training and validation data for supervised learning, along with n-gram models and utterance templates for the generator. Subsequently, the parsed data are used to perform supervised learning on the relationship between utterances and dialogue acts, resulting in a model denoted as **SL**. Finally, the pretrained SL model is reinforced with the three reward functions (Section 2.4.2), producing the models $\mathbf{RL_{utility}}$, $\mathbf{RL_{fair}}$, and $\mathbf{RL_{length}}$. Our experimental setup encompasses a total of eight models: SL and RL models using rule-based parsers and SL and RL models using deep learning-based parsers.

5.3 Evaluation Setup

Negotiation dialogue experiments were performed using our model in a web application based on previous work (He et al ., 2018). Figure 3 shows the negotiation screen of this web application. Users are presented with the scenario and item description in the upper right corner and then interact with a randomly selected model. Messages from the model appear in the box located at the bottom left, and users can input their reply messages in the chat box below. Upon reaching an agreement, users input the final offer price in the box labeled "Final Agreement" on the right. If the partner submits a final offer, users can either accept or reject it. Because the model does not consistently generate perfectly context-appropriate utterances, the dialogue may sound unnatural at times. Should users encounter difficulty continuing the dialogue, they can terminate the negotiation by pressing the "Quit" button at the bottom left.

Ten subjects were recruited for the experiment. Each subject participated in a total of 10 negotiation dialogues, with all eight models being selected with equal probability.

The experiment uses five evaluation indices. In addition to the three indices used in reinforcement learning—Utility, Fairness, and Length—we also include Agreement Rate and Human-likeness. Agreement Rate is the proportion of negotiations that resulted in an agreement, which is the primary objective in negotiation scenarios. It is calculated by dividing the number of successful agreements by the total number of dialogues performed for each model. Human-likeness is an indicator of the model's human-like behavior during dialogue. After each negotiation, users were asked, "Do you think your partner demonstrated reasonable human behavior?" They then rated the dialogue content on a 5-point Likert scale.

Figure 3: The negotiation screen on the web application.

5.4 Result

Table 7 presents the results of the human evaluation experiment for the negotiation dialogue system, grouped by optimization goal. The highest-performing results in each group are highlighted in bold. Table 8 presents an example of dialogue between a human and the negotiation dialogue system. SL(deep) demonstrated a more conscious understanding of the dialogue flow and pursued actions aligned with its own interests. As illustrated in Table 8(b), SL(deep) not only responded appropriately to the other party's price offers but also to other questions. In the context of price negotiations, it did not simply present a price but also attempted to persuade the user by providing justifications for the offered price. This characteristic is evident in the improvements observed in Utility and Fairness, as presented in Table 7. However, SL(rule) outperformed SL(deep) in terms of the negotiation agreement rate across all three optimization objectives. This is likely because Utility and Fairness scores were lower for SL(rule), leading to earlier compromise and agreement.

No considerable difference in human-likeness was observed between SL(deep) and SL(rule). Both models effectively generated dialogue involving price offers or acting as questioners. However, when acting as sellers, they struggled to respond appropriately to questions about the item's condition. Additionally, with SL(deep), the accurate determination of dialogue acts through deep learning sometimes leads to inflexibility in utterance decisions. For instance, in the initial exchange of Table 8(b), SL(deep) repeats "hello" twice. This is because that deep learning has established a pattern where an "intro" dialogue act from the other party often elicits an "intro" response. These limitations can be attributed to the search-based generator. Search- based generators exhibit limited utterance variation because all templates are derived from a training dataset. Consequently, while the dialogue act accurately captures the general structure of the utterance, the generator's capacity for detailed expression in natural language remains insufficient, resulting in no observable improvement in human-likeness.

Overall, the models were more inclined to make compromises compared to humans, indicating that they remain less effective at negotiation and highlighting areas for further improvement.

Table 7: Evaluation results of a human-assisted negotiation dialogue system (highest scores in bold).

	Ag ↑	Ut ↑	Fa ↑	Len ↑	Hu ↑	# of Examples
SL(rule)	**0.85**	-0.49	-0.98	12.31	3.4	13
SL(deep)	0.80	**-0.44**	**-0.88**	**16.40**	**3.5**	15
$RL_{utility}$(rule)	**0.55**	-0.35	-0.70	**18.91**	**2.6**	11
$RL_{utility}$(deep)	0.41	**0.23**	**-0.46**	13.25	2.4	12
RL_{fair}(rule)	**0.86**	-0.81	-1.62	13.21	**3.1**	14
RL_{fair}(deep)	0.73	**-0.57**	**-1.14**	**13.27**	2.8	11
RL_{length}(rule)	0.54	**-0.16**	**-0.32**	**15.77**	2.8	13
RL_{length}(deep)	**0.75**	-0.36	-0.72	11.92	**3.4**	12

Table 8: Example conversation between a human and a negotiation dialogue system (model utterances in bold).

Title: New EC Technology Portable Bluetooth Speaker 6W

Listing Price: $15

Description: "EC Technology® 6W Dual-Driver Portable Bluetooth STEREO Speaker w/ Built in Cell Phone/Car Speakerphones -Black", "From the manufacturer:", "Incredible sound quality surprising fullness realized through 2*3W driver, distinctive Cavity Design."

(a) A: Human (Buyer) B: SL(rule) (Seller)

B: **hi, you're interested in the item?** A: hi, i'm interested in it. how about $11?
B: **not its too low** A: hmm, how about your offer?
B: **15** A: ok, well, how about $13
B: **i think i can do 12**
A: OFFER($12) B: **ACCEPT**

Title: MODERN Contemporary White Queen Leather Bed Upholstery Tufted Italia

Listing Price: $349

Description: "This modern VERY BEAUTIFUL Queen bed is inspired by Italian design and has an Asian rubberwood construction. A white, vinyl upholstery with a tall, square-tufted headboard and clean lines highlight this elegant bed", "Included: One (1) headboard, one (1) footboard, one (1) set of rails, three (3) slats", "Materials: Asian rubberwood", "Wood finish: Black", "Upholstery materials: Vinyl, Faux L

(b) A: Human (Buyer) B: SL(deep) (Seller)

B: **hello** A: hello, i'm interested in this item.
B: **hello** A: do you have any information about this item?
B: **very small stains mostly wear and tear** A: how long have you used?
B: **3 months** A: nice. well i want to buy it in $250. what do you say?
B: **sorry there are 6 item even at 349 dollars that's a steel, less than 348 dollars a chair.**
A: how about 265? B: **lol**
A: ok, how about 275? B: **certification will be provided. 348 last price**
A: how about 300? B: **i can do 298.**
A: OFFER($12) B: **ACCEPT**

6 CONCLUSIONS

In this study, we addressed three key tasks: creating a dataset with new dialogue act proposals, developing a deep learning-based parser, and performing

dialogue experiments with a negotiation dialogue system using the deep learning-based parser.

Compared to traditional rule-based approaches, the deep learning-based parser demonstrated improved accuracy in classifying utterances into their correct dialogue acts while considerably reducing the number of utterances classified as "unknown." Among the various pretrained models evaluated, RoBERTa achieved the highest classification accuracy, while ALBERT effectively minimized the decline in accuracy while simultaneously reducing computational complexity and execution time. Moreover, in the negotiation dialogue experiments, the system using the deep learning-based parser exhibited enhanced performance in terms of utility and fairness.

This paper primarily focused on enhancing the parser component of the module framework using dialogue acts. Consequently, the performance of the dialogue act approach can be further optimized by improving the remaining managers and generators. In recent years, the LLM approach has emerged as the dominant method for chatbots (Fu et al ., 2023) (Zhao et al ., 2023). Therefore, we are currently exploring the integration of LLMs as generators, incorporating dialogue acts into the prompts (Wagner et al ., 2024). Furthermore, we will perform dialogue experiments comparing our proposed method with LLMs, aiming to further demonstrate the value of the dialogue act approach, which effectively captures the structural outline of an utterance.

REFERENCES

Asher, N., Hunter, J., Morey, M., Benamara, F., & Afantenos, S. (2016). Discourse Structure and Dialogue Acts in Multiparty Dialogue: the STAC Corpus, In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pp. 2721-2727.

Cano-Basave, A. E., & He, Y. (2016). A study of the impact of persuasive argumentation in political debates. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages pp.1405-1413.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (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 (NAACL-HLT)*, p.4171-4186.

Fisher, R., Ury, W. L., & Patton, B. (2011). Getting to yes: Negotiating agreement without giving in. Penguin.

Fu, Y., Peng, H., Khot, T., & Lapata, M. (2023). Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv:2305.10142.*

He, H., Chen, D., Balakrishnan, A., & Liang, P. (2018). Decoupling Strategy and Generation in Negotiation Dialogues, In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2333-2343.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite Bert for Self-supervised Learning of Language Representations, In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Lewicki, R. J., Saunders, D. M., & Minton, J. M. (2011). Essentials of negotiation. McGraw-Hill/Irwin Boston, MA, USA.

Lewis, M., Yarats, D., Dauphin, Y. N., Parikh, D., & Batra D. (2017). Deal or No Deal? End-to-End Learning for Negotiation Dialogues, In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2443-2453.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019) .Roberta: A Robustly Optimized BERT Pretraining Approach, *arXiv:1907.11692*.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, *arXiv:1910.01108*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need, In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 6000-6010.

Wagner, N., & Ultes, S. (2024). On the Controllability of Large Language Models for Dialogue Interaction. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 216-221.

Young, S., Gašić, M., Thomson, B., & Williams, J. D. (2013). POMDP-Based Statistical Spoken Dialog Systems, In A review. In Proceedings of the IEEE, Vol.101(5), pp. 1600-1179.

Żelasko, P., Pappagari, R., & Dehak, N. (2021). What Helps Transformers Recognize Conversational Structure? Importance of Context, Punctuation, and Labels in Dialog Act Recognition, *Transactions of the Association for Computational Linguistics*, Vol.9, pp.1163-1179.

Zhan, H., Wang, Y., Feng, T., Hua, Y., Sharma, S., Li, Z., Qu, L., & Haffari, G. (2020). Let's Negotiate! A Survey of Negotiation Dialogue Systems. *arXiv:2212.09072.*

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., et al. (2023). A survey of large language models. *arXiv:2303.18223.*