

Would You Like to Make a Donation? A Dialogue System to Persuade You to Donate

Yuhan Song*, Houfeng Wang*,†

*National Key Laboratory of Multimedia Information Processing
School of Computer Science, Peking University
{songyuhan, wanghf}@pku.edu.cn

Abstract

Persuasive dialogue is a type of dialogue commonly used in human daily life in scenarios such as promotion and sales. Its purpose is to influence the decision, attitude or behavior of another person through the dialogue process. Persuasive automated dialogue systems can be applied in a variety of fields such as charity, business, education, and healthcare. Regardless of their amazing abilities, Large Language Models (LLMs) such as ChatGPT still have limitations in persuasion. There is few research dedicated to persuasive dialogue in the current research of automated dialogue systems. In this paper, we introduce a persuasive automated dialogue system. In the system, a context-aware persuasion strategy selection module makes dialogue system flexibly use different persuasion strategies to persuade users; Then a natural language generation module is used to output a response. We also propose a persuasiveness prediction model to automatically evaluate the persuasiveness of generated text. Experimental results show that our dialogue system can achieve better performance on several automated evaluation metrics than baseline models.

Keywords: persuasion, dialogue systems, language models

1. Introduction

Automated dialogue system is a computer program used to simulate human dialogue. It is an important research direction in natural language processing of artificial intelligence. It has a wide range of applications, such as chatbots, customer consultation, intelligent assistants, etc., which can not only provide people with convenient and fast interaction, but also improve the efficiency of service and user experience. With the progress of deep learning, pre-trained language models and recent large language models, automated dialogue systems based on a large number of training data and network parameters can flexibly respond to more dialogue scenes and generate more smooth dialogue, and are increasingly applied in various fields.

Persuasive dialogue is a common kind of dialogue in human daily life, which aims to make another person accept a point of view or do something. Such dialogue usually occurs in publicity, sales and other scenarios. However, even they show strong performance in natural language generation tasks and attract much attention recently, the large language models such as ChatGPT still have shortcomings when it comes to persuasive dialogue. As shown in Table 1, ChatGPT usually only gives macro arguments, which are related to society and the world, when persuading users to donate to a charity organization, while human's persuasive utterance has more specific and person-

alized expressions, and can also make use of the foot-in-the-door effect in psychology for persuasion so as to achieve better persuasive effects.

Persuader	Utterance
ChatGPT	Your donation can help to support children who are living in some of the most challenging and dangerous environments in the world. From war-torn countries to impoverished communities, Save the Children works tirelessly to ensure that every child has a chance to reach their full potential.
Human-1	I guess there are more people than we know who grew up with worries. I grew up worrying about my safety and whether I'd have a place to live. My dad was a gambling addict. So I think the work that charities can do is so important.
Human-2	You can give any amount, even as little as \$0.01 can help.

Table 1: A contrast between ChatGPT and human persuasion utterances. The human utterances come from the PersuasionForGood dataset. See Section 3 for more details.

Persuasive automated dialogue systems can be used in many situations. We can make the system persuade users to accept a certain point of view. For example, we can promote users' physi-

†Corresponding author.

cal exercise and healthy diet through persuasion of the model (Zhang et al., 2020). For society, we can also make the system try to persuade users to donate to charity, do what they can to protect the environment, and so on. Therefore, the study of such dialogue systems with persuasive functions can achieve the purpose of technology for social good.

In this work, we propose a persuasive dialogue system. Instead of training a generative model on a persuasive dialogue dataset, we consider selecting a persuasion strategy first and then explicitly input the chosen strategy into the generative model by using additional special tokens. Therefore, the generative model can learn the style of utterance with different persuasion strategy and perform better in generation task when it is given a certain strategy. Based on the above considerations, our model is composed of a persuasion strategy selection module and a natural language generation module. Moreover, the dialogue we are considering is not single-round. The dialogues usually last for several turns, and in each turn the participants may speak several sentences. So it is essential to make the model aware of the dialogue context. For the persuasion strategy selection module, a vector representation of the dialogue context is input to help the module make a better choice; For the natural language generation module, we input all the history dialogue and let it generate a response accordingly.

Usually, the evaluation of a dialogue system includes the relevance and the diversity aspects. But when it comes to persuasive dialogue system, the persuasiveness matters a lot. Since manual evaluation is time-consuming and costly, we propose a persuasiveness prediction model which can quantify the persuasiveness of dialogue. We then use a trained persuasiveness prediction model to evaluate our system in terms of persuasiveness. Experiments show that our proposed persuasive dialogue system outperforms several baseline models, and its performance is close to the human performance in the dataset.

The contributions of this paper are as follows¹:

- We propose a persuasive automated dialogue system, supplementing the research on this kind of dialogue system.
- We implement the proposed system based on PersuasionForGood dataset, and the system can persuade the user to donate to a certain charity organization, which practice the slogan of technology for social good.
- We also propose a persuasiveness prediction

model, promoting automated evaluation of dialogue persuasiveness. It can replace time-consuming and costly manual evaluation to some extent.

2. Related Work

The development of dialogue systems has a long history. The earliest generative dialogue system is ELIZA (Weizenbaum, 1966), which is based on rules. Nowadays most of the dialogue systems are based neural networks. The first to appear and receive much attention is the sequence-to-sequence model based on Recurrent neural networks (RNN) (Sutskever et al., 2014). Then came the sequence-to-sequence model based on Long Short Term Memory (LSTM) (Vinyals and Le, 2015). A hierarchical recursive encoder-decoder model with latent variables VHRED (Serban et al., 2017) was also proposed for open domain dialogue generation, which uses random variables to model the hierarchy of dialogue sequences and input them into RNN. The Transformer-based generation system includes a conditional Transformer language model (Keskar et al., 2019) and Meena (Adiwardana et al., 2020). There are also many pre-trained generative language models based on Transformer, such as the GPT series model (Radford et al., 2019; Brown et al., 2020; OpenAI, 2023) and T5 (Raffel et al., 2020).

At present, research on persuasive dialogues is still few. Related studies include datasets and baseline work on persuasive dialogues, analysis of persuasive strategies and persuasive effects, research on resistance strategies of the persuadee, and research on the persuasiveness of utterance, etc. The directions of existing research are diverse.

Some traditional persuasive automated dialogue systems are used in the fields of law (Gordon, 1993), car sales (André et al., 2000), intelligent tutoring (Yuan et al., 2008), etc., but none of them have applied deep learning models. In recent years, there is work that studied persuasive conversations on topics of current social events (Lukin et al., 2017), but it only considered single-round conversations.

Wang et al. (2019) collected and released PersuasionForGood, a dataset of real human persuasive conversations, where the goal of the dialogue is to persuade another person to donate money to charities. The dataset has detailed annotations on persuasion strategies, psychological characteristics of both parties of the conversation, and the final amount of donation. Accordingly, they trained a Recurrent-CNN model for classifying the persuasion strategy used in the input text. Finally, based on the dataset, they analyzed in detail the associations between participants' donation behavior and

¹Our code are released at <https://github.com/QbethQ/PersuasiveDialogue>.

their sociodemographic backgrounds, personality traits, and value systems, as well as the differences in the persuasive effects of different persuasion strategies on people with different psychological profiles.

Some subsequent work is also based on the PersuasionForGood dataset. Tian (2022) investigated the resistance strategies of the persuadee. The author labeled the resistance strategies in the dataset, and similarly trained a Recurrent-CNN model for classifying the resistance strategies used in the input text. In addition, this work analyzed the impact of these resistance strategies on the persuadee’s donation behavior and the effectiveness of different persuasion strategies in the face of different resistance strategies. There is also work that investigated the role of emotions in persuasive dialogues (Peng et al., 2023). They proposed a model that can sense the emotions of a conversation, and conducted experiments on the PersuasionForGood dataset.

There is also dataset with persuasive dialogues for task-oriented dialogue systems. Raut et al. (2023) collected and released the personalized persuasive dialogue (PPD) dataset. It includes a large number of conversations in the electronic commerce scenario, in which salesmen try to persuade customers to buy a product. This work also proposes an end-to-end persuasive task-oriented dialogue system based on reinforcement learning and meta-learning on this dataset.

In addition, there is also work that investigated and modeled the persuasiveness of text. Hidey and McKeown (2018) trained a model that can compare the persuasiveness of two span of text, using data from persuasive conversations on the online forum Change My View; Yang et al. (2019) proposed a neural network that can quantify the persuasiveness of an article and identify the persuasion strategy of each sentence. However, their work only focused on single-turn dialogues and long articles respectively, instead of multi-turn dialogues.

3. The PersuasionForGood Dataset

We perform our experiments and analysis on the PersuasionForGood dataset (Wang et al., 2019), which was collected from online human-human conversations on an persuasion task. In each conversation, one participant (persuader) needed to persuade the other participant (persuadee) to donate some money to *Save the Children*², which is one of the most well-known charity organizations all around the world. The persuaders were provided with tips on different persuasion strategies with some example sentences. Therefore, the

²<https://www.savethechildren.org/>

persuaders could use different persuasion strategies to achieve the goal. In the dataset, there are 10 types of persuasion strategies and one "None" type. Each utterance from the persuaders in the 300 annotated dialogues was manually labeled with a persuasion strategy. Moreover, the amount of the persuadee’s donation after the dialogue was recorded. There are also some other information in the dataset about the participants, such as education backgrounds, income, and their psychological feature vectors generated by psychological questionnaires.

4. Method

Just like human’s behavior at the time of data collection, we also make our model select a persuasion strategy and then generate a response. The proposed model consists of a persuasion strategy selection module and a natural language generation module. The overview of our model is shown in Figure 1.

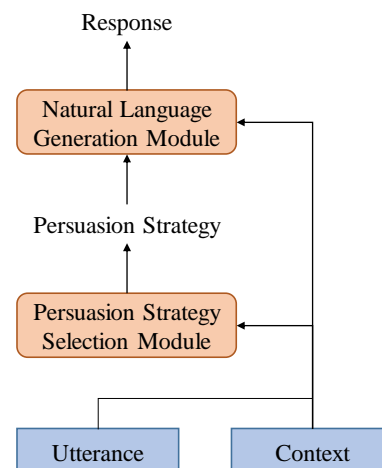


Figure 1: The overview of proposed persuasive dialogue system. The user utterance and the conversation context are input into the persuasion strategy selection module, which will output a persuasion strategy. Then the natural language generation module takes the user utterance, the conversation context and the persuasion strategy as input to generate a response.

In addition to traditional automated evaluation metrics of dialogue systems, we also hope to automatically evaluate the persuasiveness of the model. Therefore, we propose a model that can predict the persuasiveness of a dialogue, i.e. the amount the persuadee is willing to donate to the charity organization mentioned in the PersuasionForGood dataset. We will describe the persuasiveness prediction model in Section 4.3.

Strategy	Utterances
Credibility appeal	Persuadee: What are some of the things that they do for the children?
	Persuader: They help to provide basic health care, provide disaster relief, and help to ensure their rights to education and safety.
Foot-in-the-door	Persuadee: I would like to help in the future when I am more financially stable.
	Persuader: I understand but even the smallest amount would be a BIG help.
Emotional appeal	Persuadee: At this time I am not willing to donate.
	Persuader: Well one statistic that in 2018 alone over 1,000 children were killed or injured in Syria. This charity’s goal is to help them out.

Table 2: Some examples of different persuasion strategies

4.1. Persuasion Strategy Selection

Persuasion Strategy Selection is a subjective and diverse task, which has a lot to do with the personality of the persuader and the persuadee and the context of the conversation. Table 2 shows some examples of different persuasion strategies. In the first sample, the persuadee had questions about what the charity organization did, and the persuader provided some information about the organization to persuade by credibility appeal. In the second sample, the persuadee did not want to donate because of financial instability, and the persuader used the Foot-in-the-door strategy to persuade. In the last sample, the persuadee expressed unwillingness to donate, and then the persuader tried to emotionally persuade by talking about some children’s tragic experiences.

In order to make the model use different persuasion strategies flexibly, we propose a context-aware persuasion strategy selection module. It selects a persuasion strategy according to the user’s current utterance and the dialogue context. Since the distribution of different strategies in various dialogue rounds is significantly different (Wang et al., 2019), the module also takes the number of current dialogue turns as input. Our persuasion strategy selection module is based on BERT (Devlin et al., 2018), which is based on Transformer (Vaswani et al., 2017). By using lots of text data for pre-training on two tasks, Masked Language Model and Next Sentence Prediction, BERT can calculate the vector representation of each word in context and the vector representation of the whole sentence. These vector representations are often used as inputs for downstream tasks, and the pre-trained BERT model can also be fine-tuned to fit the needs of a particular task. Because of BERT’s bidirectional encoding of context, it achieved the best performance at that time on many downstream tasks, especially on natural language understanding tasks such as text classification. The persuasion strategy selection module’s architecture is shown in Figure 2.

We can formulate the module. Let $v_s = \text{BERT}(s)$ represent the sentence-level vector representation of sentence s . Concretely, We use the pooled out-

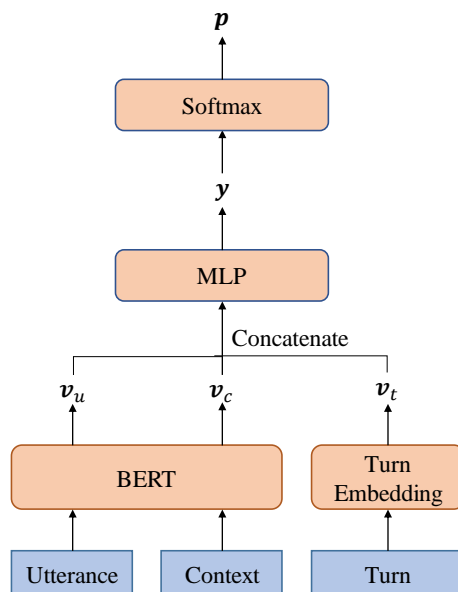


Figure 2: The architecture of the persuasion strategy selection module.

put, i.e. the output of passing the CLS token output into a dense layer and a Tanh activation function, as the sentence-level vector representation. When the user’s t^{th} utterance u_t is input into the module, the conversation history $H_t = \{s_1, u_1, \dots, s_t\}$ has been stored in the module’s memory, where s_i represents the i^{th} utterance of the dialogue system. Then the module calculate the vector representation of the current user utterance and the context:

$$v_u = \text{BERT}(u_t), v_c = \text{BERT}(s_1 || u_1 || \dots || s_t), \quad (1)$$

where the symbol $||$ represents concatenating operation. And the current dialogue turn t is input into an embedding layer to generate a turn embedding vector v_t :

$$v_t = \text{TurnEmbedding}(t). \quad (2)$$

Then we concatenate the above 3 vectors as the input of multilayer perceptron (MLP) to get the logits:

$$y = \text{MLP}(v_u || v_c || v_t). \quad (3)$$

The final probability distribution of each strategy is obtained by passing the logits y through a Softmax layer:

$$p = \text{Softmax}(y). \quad (4)$$

We attempt two kinds of classifier settings. One is to use cross-entropy loss to train a single label classification model; the other one is to use binary cross-entropy loss to train a multi-label classification model to find all strategies for one turn, since the persuader may speak several sentences with different persuasion strategies in each turn. The experiment results and our analysis are presented in Section 5.2.

4.2. Natural Language Generation

Generative Pre-trained Transformer 2 (GPT-2) (Radford et al., 2019) is also a Pre-trained Language Model based on Transformer (Vaswani et al., 2017). Unlike BERT, GPT-2 uses a unidirectional language model and consists of decoders in a multi-layer Transformer structure. It performs self-supervised learning based on large amounts of data, training the model to constantly predict the next word in the input text, thus learning the ability of context representation and text generation. GPT-2 has demonstrated its powerful performance in many text generation tasks, such as dialogue generation, machine translation, and question answering.

Due to its powerful ability to produce natural, smooth, coherent and high-quality text, and its strong generalization ability. We will use the GPT-2 model fine-tuned on the PersuasionForGood dataset for the natural language generation module, which is part of the proposed persuasive dialogue system.

To adapt the GPT-2 model to natural language generation task on the PersuasionForGood dataset, we added some special tokens to the vocabulary to make the model aware of the speaker of the current conversation, and the persuasion strategy used by the utterance. Specifically, as shown in Table 3, we add two tokens <persuader> and <persuadee> before each sentence to indicate the speaker of the following utterance to help the model understand the dialogue. Before each sentence of the persuader, if the utterance uses a certain persuasion strategy, We also specify the name of the strategy between the special tokens <policy_begin> and <policy_end>. These newly added special tokens' word embedding vectors will be trained during the fine-tuning phase, and they will help the model better understand the dialogue and learn the text paradigm of each persuasion strategy, and

enable it to generate the corresponding response with each persuasion strategy. In addition, in the self-supervised training of unidirectional language model, only the utterance spoken by the persuader is what the model needs to learn to generate, and other information only helps the model understand the dialogue context of the generation task.

4.3. Persuasiveness Prediction

To evaluate the model's capabilities comprehensively, we propose a persuasiveness prediction model. The model takes the dialogue as input and predicts the amount of the persuadee's donation after the conversation. Figure 3 shows the overview of the persuasiveness prediction model.

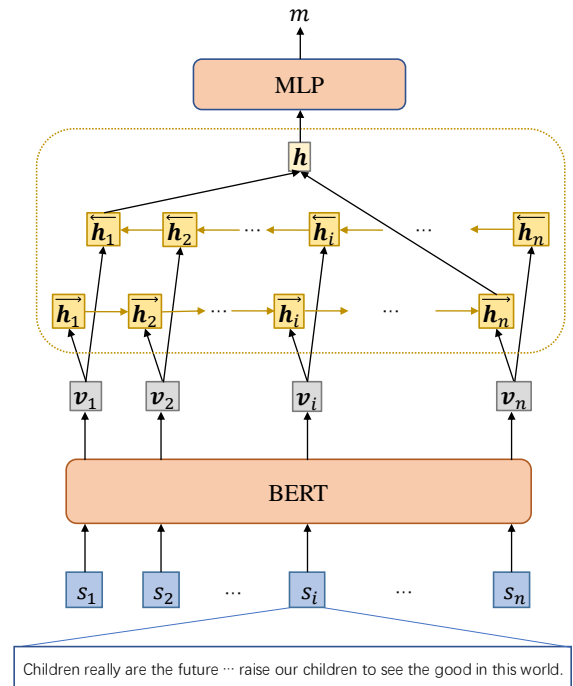


Figure 3: The overview of the persuasiveness prediction model. The part in the dotted box is a BiLSTM.

Given a n -round dialogue

$$D = \{s_1, u_1, \dots, s_{n-1}, u_{n-1}, s_n\}$$

where s_i, u_i represent the i^{th} utterance of the dialogue system and that of the user respectively. Since our dialogue system only plays the role of the persuader and we cannot get real user's utterances, we only take all the persuader's utterances as input. The model first calculates the sentence-level vector representations:

$$v_i = \text{BERT}(s_i), \text{ where } i = 1, 2, \dots, n. \quad (5)$$

Since BERT only encodes information from every sentence, while we need to model the information

Text	Persuadee: I am not sure whether to donate or not. Persuader: I read an article yesterday that said 1000 children were saved from hunger shelters just by donations from online workers.
Tokens	<persuadee> I am not sure whether to donate or not. <SEP> <persuader> <policy_begin> emotion -appeal <policy_end> I read an article yesterday that said 1000 children were saved from hunger shelters just by donations from online workers. <SEP>

Table 3: A span of dialogue text with its corresponding token sequence in the GPT-2 model.

of the whole dialogue, we choose to use bidirectional Long Short Term Memory network (BiLSTM) to bidirectionally encode information among the utterances. Formally, a subsequent BiLSTM calculates a dialogue vector representation:

$$\vec{h}_i = \text{LSTM}_{\text{forward}}(v_i, \vec{h}_{i-1}), \quad (6)$$

$$\overleftarrow{h}_i = \text{LSTM}_{\text{backward}}(v_i, \overleftarrow{h}_{i+1}), \quad (7)$$

where $i = 1, 2, \dots, n$.

We use the concatenation of the final hidden state of each direction as a high-level representation of all the persuader’s utterances, and it is used as features that are input into a MLP for predicting m , the amount of the persuadee’s donation, which may stand for the persuasiveness to some extent:

$$m = \text{MLP}(\vec{h}_n \parallel \overleftarrow{h}_1). \quad (8)$$

We apply the mean squared error between the predicted and ground truth amount of donation as training loss and the final evaluation metric of the persuasiveness prediction model:

$$l = (m_{\text{predict}} - m_{\text{truth}})^2. \quad (9)$$

5. Experiment

5.1. Settings

In this paper, we use the pre-trained BERT (base, uncased) model and the corresponding tokenizer released by Huggingface³ in the persuasion strategy selection module and the persuasiveness prediction module. We freeze the parameters of BERT in our experiment. And the pre-trained GPT2 model on English language is used in the natural language generation module⁴. We use the Adam (Kingma and Ba, 2014) optimizer with the learning rate of $1e-4$.

There are 300 annotated dialogues in the PersuasionForGood dataset and we divide them into training set, validation set and test set in a ratio of 7:1:2. We randomly initialize the parameters except those of the pre-trained language models. The persuasion strategy selection module and the

persuasiveness prediction model are trained for 30 epochs, and the natural language generation module for 50 epochs. Then we choose the model that performs the best on the validation set.

5.2. Results

5.2.1. persuasion strategy Selection

The experimental results for the persuasion strategy selection module on PersuasionForGood dataset are shown in Table 4.

Model Setting	Accuracy
Single Label Classification	46.40%
Multi Label Classification	31.79%

Table 4: The experimental results of the persuasion strategy selection module.

The accuracy of multi-label classification is lower, which is due to the fact that multi-label classification task is more difficult to accurately selecting all persuasive strategies consistent with the ground truth label for a turn of dialogue. As mentioned earlier, persuasion strategy selection is a subjective and diverse task, and there may be not a fixed standard answer. Therefore the model’s low accuracy in persuasion strategy selection does not indicate that the model’s performance is poor; on the contrary, the model could select persuasive strategies that are different from the ground truth label in the dataset in some cases, providing a high degree of flexibility and variety. In addition, the model is able to achieve an accuracy rate close to 50% on a ten-class classification task, which is also sufficient to cope with some situations where the persuasion strategy should be fixed, such as answering donation information when the user asks a question. The system that combines the persuasion strategy selection module with the natural language generation module also shows good performance (see the case study in Section 6). In general, the persuasion strategy selection module shows satisfactory performance.

5.2.2. Persuasiveness Prediction

For the persuasiveness prediction model, the distribution of the amount of donation and our experi-

³<https://huggingface.co/bert-base-uncased>

⁴<https://huggingface.co/gpt2>

Data	Mean	MSE (variance)	Pearson correlation coefficient
Raw Data	0.2264	0.1663	-
Full Model Prediction	0.2357	0.1761	0.31
Model Without BiLSTM Prediction	0.1547	0.2294	0.20

Table 5: The experimental results of the persuasiveness prediction model. We list the raw data distribution, the Mean Squared Error (MSE) of the model prediction (for raw data, the variance), and the Pearson correlation coefficient between the ground truth and the prediction of our persuasiveness prediction model on test set. Here we exclude the data points with donations greater than \$5.

Model	Relevance			Diversity		Persuasiveness
	BLEU-4	Average	IDK	Distinct-1	Distinct-2	Predicted Donation
MMI-bidi	1.72	0.787	44.4%	4.73	16.26	0.1783
kg-CVAE	1.78	0.834	47.6%	4.99	18.24	0.2016
CT	1.96	0.786	58.8%	6.60	27.46	0.1841
Finetuned GPT-2	2.38	0.783	91.1%	10.10	42.41	0.1953
Ours	2.82	0.814	100%	10.43	44.86	0.2276
Human	-	-	-	11.16	50.54	0.2357

Table 6: The evaluation results of the responses generated by our whole system. "Average" is short for Word Embedding Average Similarity. "Human" represents human conversations in the dataset. The value in bold indicates the best performance under corresponding metric (except for Human).

mental results are listed in Table 5. What calls for special attention is that the amount of donation may appear some abnormal data point. For example, someone with a huge wealth and a strong willingness to donate could give as much as \$100, while others donated less than \$1 on average. Considering that the MSE loss function is very sensible to such abnormal data, we exclude all the data points with donations greater than \$5, which account for about only 3% of the dataset.

Our proposed persuasiveness prediction model achieves an average prediction and a mean squared error that are both close to the raw data mean and variance. Even though we exclude some extreme data points, there are still some which deviate from the average a lot. Under the influence of these data points, our proposed model shows good performance as well.

For ablation study, we also train and test a persuasiveness prediction model without the BiLSTM layer. In the model without BiLSTM, the output of BERT is directly input into the MLP to calculate the predicted amount of donation. The results are also listed in Table 5. We can conclude that the prediction of model without BiLSTM deviates farther from the ground truth value, and shows a more different distribution, causing a much lower correlation coefficient.

5.2.3. The Whole System

Finally we are going to evaluate our whole proposed persuasive dialogue system. We compare our model with the following baseline models from previous work: (1) **MMI-bidi**: a sequence-to-sequence model with response re-ranking trained by maxi-

mizing the mutual information (Li et al., 2015); (2) **kg-CVAE**: a conditional variational autoencoders based on knowledge and dialogue act (Zhao et al., 2017); and (3) **CT**: a sequence-to-sequence model whose decoder is conditionally trained by feeding the embedding of pre-defined response attributes (See et al., 2019). To show the effect of the persuasion strategy selection module, we also compare the model with a GPT-2 model which is fine-tuned on the dataset.

We evaluate these models from the perspective of relevance, diversity and persuasiveness. In terms of the diversity of the text, the generated text was also compared with human utterances in the dataset. For the metrics indicating the relevance of the generated text to the reference text in the dataset, we calculate the BLEU-4 (Papineni et al., 2002) score and Word Embedding Average Similarity (Average) (Mitchell and Lapata, 2008). However, these two metrics depend too much on the ground truth response, and may correlate poorly with dialogue relevance. So we also use the IDK (Berlot-Attwell and Rudzicz, 2022) metrics to evaluate our proposed system. In order to facilitate comparison, we normalized the IDK results to make that of our proposed system 100%, and present the results of baseline models as percentages. For the metrics indicating the diversity of the generated text, we follow Li et al. (2015) to use Distinct-1 and Distinct-2, i.e., the ratios of distinct unigrams and bigrams in all generated texts. The experimental results are shown in Table 6.

The results of the BLEU-4 metric for our proposed natural language generation module exceeds those of other models, and the IDK metric results also shows that our proposed system achieves

System	Hi! How are you today?
User	Hi! I'm fine, and you?
System	Fine! I'm helping to raise money for an organization called Save the Children.
User	Oh, what about this organization?
System	I was just reading about them. They work in 120 countries and are connected to big companies.
User	That sounds good. What do they do for children?
System	For instance, millions of Syrian children have grown up facing the daily threat of violence. over 1,000 children have been reportedly killed or injured in intensifying violence. Donations are needed to ensure children's rights to health, education, safety, etc.
User	It's really terrible.
System	So do I have your commitment to the organization? I'm really pleased to hear that your donation is essential to ensure children's rights to health, education, safety, etc. If you can't donate now, why not now?
User	I'm willing to donate half of my rewards!
System	That's fantastic - \$2 would be so much more! You can also donate in increments from your reward. You can choose any amount!
User	OK. How will the money be used?
System	The research team will collect all the donations and send it to Save the Children. It will go to pay for flights to and warehousing for the kids and send them to schools all over the world! The researchers also donate on request to charity.
User	That's really good.
System	Thank you so much! You and I in the meantime! The research team will send the organization your donation today!
User	OK, that's fine.
System	I'm really pleased that you are willing to donate your rewards today. Thank you again for your generous donation and I hope you have a wonderful day!

Table 7: A sample dialogue of the proposed persuasive dialogue system

better performance in terms of relevance of the generated text, though there is a small gap in terms of word embedding average similarity between the proposed module and the kg-CVAE model. In addition to the significant improvement in terms of relevance that GPT-2 model brings, our proposed persuasion strategy selection module also contributes to this. Our module also generates text with far better diversity than all the baseline models, and is quite close to the human utterances in the dataset. Moreover, according to our persuasiveness prediction model, the dialogue generated by our proposed system might result in more donations, an amount close to that caused by the human's utterances.

6. Case Study

By combining the persuasion strategy selection module with the natural language generation module, we can implement an interactive automated persuasive dialogue system. An example of the dialogue it generates is shown in Table 7.

After greetings, the system goes straight to the topic and starts talking and introducing the charity organization Save the Children. Then, the system talks about the plight of Syrian children to explain the importance of donations through logical appeals strategy. After that, through emotional appeal strategy, it expresses the necessity of the user's donation and makes a donation request to the user.

After the user has expressed a willingness to donate, the system also uses the foot-in-the-door effect to hope that the user will increase the amount of donation. When the user asks how the donated money will be used, the system persuades users with credibility appeals strategy through the facts related to the charity organization. Finally, the system expresses thanks twice and ends the conversation. The system not only completes the task of persuasion, but also generates relatively smooth and fluent sentences. The selection of persuasion strategies is also suitable and diversified, indicating the good performance of the system.

There are still some shortcomings in the text generated by the system. When the user asks what the charity organization could do for children, the system only responds to the plight of children in Syria and explains the importance of donations to children's health, education and safety, instead of directly answering the user's question. There are also a small number of incoherent sentences in the generated text, such as "If you can't donate now, why not now?" when making a donation request to the user. In this sentence, the semantics of the main sentence and the following rhetorical question are inconsistent. There are also some repetition in the generated text, such as the system repeating "the research team will send the donation to the charity organization" twice in the third and the second turn of the dialogue from bottom. These

drawbacks need to be solved by further research, such as making the model pay more attention to the context and introducing penalties for the repeated text generated by the model.

7. Ethical Considerations

Persuasion, as a fundamental aspect of human interaction, carries both advantages and disadvantages. On the positive side, persuasion can be a powerful tool for promoting understanding, cooperation, and people's beneficial behaviors.

However, persuasion also has its drawbacks. It can be used manipulatively, leading to the spread of misinformation or evil use. Thus, while persuasion can bring positive changes when used ethically, its misuse poses significant risks to social well-being.

Due to the rapid development of dialogue systems, we must take ethical principles into consideration through all stages of designing a persuasive dialogue system. Particular attention should be paid to the training dataset because it decides the intention of a system. Therefore, we choose to use a public persuasive dialogue dataset on donation task, which is beneficial for children and society.

Moreover, we should be cautious about using persuasive dialogue systems in practice. Firstly, There should be ongoing oversight to ensure that the models do not output false or even toxic content. Secondly, the users, i.e. the persuadees, should know in advance that these content is generated by models. Thirdly, some further persuasive dialogue system may generate responses according to the user's psychological profiles and sociodemographic backgrounds. The collection of these information should require the consent of the user and should be strictly regulated.

In short, from training data to model deployment, there should be continuous attention and monitoring to ensure that the persuasive dialogue system generates true and harmless output, and the purpose of persuasion should be consistent with legal and ethical standards.

8. Conclusions and Future Work

In this paper, we conduct a research on persuasive dialogue and implement an interactive automated persuasive dialogue system based on the PersuasionForGood dataset. Compared with those of related work, experimental results show that the proposed persuasive automated dialogue system has better performance than several baseline models. In the case study, we can also see that the model can complete the task of persuasion well through the flexible use of different persuasion strategies. Our proposed persuasiveness prediction model achieves good performance as well.

Moreover, We also propose a persuasiveness prediction model for automated evaluation of the persuasiveness of generated text. It has shown good performance in prediction the amount of donation. It also shows that the persuasiveness of our persuasive dialogue system outperforms the baseline models. The proposal of the persuasiveness prediction model promoted the automation of the evaluation of the persuasiveness of generated text, which is good for research on persuasive dialogue systems.

However, the automated persuasive dialogue system proposed in this paper only considers the context and the persuasion strategies, but does not explicitly consider the user's resistance strategy and the role of emotion in the dialogue. In terms of dialogue generation, the model still has a few problems of semantic incoherence and insufficient attention to context, and the responses generated by the model sometimes repeat themselves. Although the model in this paper shows good performance on many automated evaluation metrics, it still lacks manual evaluation which is more reliable. In addition, from the design of the whole system, the model does not have the ability to extract the user's personality characteristics from the dialogue and flexibly adjusting the dialogue and persuasion strategies accordingly, which is necessary for the personalized and targeted persuasion.

Currently, there are few high-quality datasets containing persuasive dialogues. In the future, high-quality datasets of persuasive dialogue can be constructed in many fields such as advertising, sales, publicity, education and counseling, and social affairs, and corresponding automated persuasive dialogue systems can be established. At that time, how to build a common persuasive dialogue system in various fields will also be one of the topics to be studied.

In addition, the evaluation methods of automated persuasive dialogue system are limited. More work can be done in the future to model and evaluate the persuasiveness of texts either qualitatively or quantitatively.

In short, in the field of automated persuasive dialogue system, the future work can be started from the construction of datasets, the design and training of models, and the evaluation of systems. This is a research direction with much space, high research significance, but also full of challenges.

9. Acknowledgement

This work was supported by National Natural Science Foundation of China (62161160339), and National Science and Technology Major Project (2022ZD0116308). The corresponding author is Houfeng Wang.

10. Bibliographical References

- Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.
- Elisabeth André, Thomas Rist, Susanne Van Mulken, Martin Klesen, and Stefan Baldes. 2000. The automated design of believable dialogues for animated presentation teams. *Embodied conversational agents*, pages 220–255.
- Ian Berlot-Attwell and Frank Rudzicz. 2022. Relevance in dialogue: Is less more? an empirical comparison of existing metrics, and a novel simple metric. *arXiv preprint arXiv:2206.01823*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Thomas F Gordon. 1993. The pleadings game: An exercise in computational dialectics. *Artificial intelligence and law*, 2:239–292.
- Christopher Hidey and Kathleen McKeown. 2018. Persuasive influence detection: The role of argument sequencing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.
- Stephanie M Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. *arXiv preprint arXiv:1708.09085*.
- Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *proceedings of ACL-08: HLT*, pages 236–244.
- OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, and Yajing Sun. 2023. Do you know my emotion? emotion-aware strategy recognition towards a persuasive dialogue system. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part II*, pages 724–739. Springer.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.
- Aritra Raut, Abhisek Tiwari, Subrata Das, Sriparna Saha, Anutosh Maitra, Roshni Ramnani, and Shubhashis Sengupta. 2023. Reinforcing personalized persuasion in task-oriented virtual sales assistant. *Plos one*, 18(1):e0275750.
- Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654*.
- Iulian Serban, Alessandro Sordani, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.
- Youzhi Tian. 2022. *Understanding user resistance strategies in persuasive conversations*. University of California, Davis.

- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.
- Xuwei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649.
- Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.
- Can Xu, Wei Wu, Chongyang Tao, Huang Hu, Matt Schuerman, and Ying Wang. 2019. Neural response generation with meta-words. *arXiv preprint arXiv:1906.06050*.
- Diyi Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. 2019. Let’s make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. 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 3620–3630.
- Tangming Yuan, David Moore, and Alec Grierson. 2008. A human-computer dialogue system for educational debate: A computational dialectics approach. *International Journal of Artificial Intelligence in Education*, 18(1):3–26.
- Jingwen Zhang, Yoo Jung Oh, Patrick Lange, Zhou Yu, and Yoshimi Fukuoka. 2020. Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet. *Journal of medical Internet research*, 22(9):e22845.
- Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2018. Learning to control the specificity in neural response generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1108–1117.
- Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*.