Persuasive Dialogue Understanding: the Baselines and Negative Results

Hui Chen[a], Deepanway Ghosal[a], Navonil Majumder[a], Amir Hussain[b], Soujanya Poria[1,*]

[a]*Information Systems Technology and Design, Singapore University of Technology and Design, Singapore*
[b]*School of Computing, Edinburgh Napier University, UK*

Abstract

Persuasion aims at changing one's opinion and action via a series of persuasive messages containing persuader's strategies. Due to its potential application in persuasive dialogue systems, the task of persuasion strategy recognition has gained much attention lately. Previous methods on user intent recognition adopt recurrent neural network (RNN) and convolutional neural network (CNN) to model context in conversational history, neglecting the tactic history and intra-speaker relation. In this paper, we propose a transformer-based framework coupled with Conditional Random Field (CRF) for persuasion strategy recognition, where we leverage inter-speaker, intra-speaker contextual semantic features, and label dependencies to improve the recognition. On the benchmark dataset, our model outperforms the strong baselines by a significant margin.

Keywords: Persuasive Dialog Systems, Transformer-based Neural Networks, Conditional Random Field, Persuasion Strategy Recognition

1. Introduction

Persuasive dialogue is an active area of research in the field of dialogue systems and is getting more and more traction recently. In a persuasive dialogue, there are two roles, a persuader and a persuadee. The persuader aims to change the persuadee's opinion

*Corresponding author.
Email addresses: hui_chen@mymail.sutd.edu.sg (Hui Chen), deepanway_ghosal@mymail.sutd.edu.sg (Deepanway Ghosal), n.majumder.2009@gmail.com (Navonil Majumder), A.Hussain@napier.ac.uk (Amir Hussain), sporia@sutd.edu.sg (Soujanya Poria)

and reach an intent by using conversational strategies. However, most previous work on persuasiveness mining mainly focuses on the detection and prediction of argumentative features [1, 2], syntactic features [3] and semantic types of argument components [4] in online persuasive forums. It has yet to receive further research on persuasiveness recognition in dialogues.

Generally, in a persuasion strategy recognition task, each utterance is accompanied by a semantic label containing the speaker's strategy, and then the goal is to identify these strategies by referring to contextual utterances. So this task can be regarded as a sequence labeling task. Table 1 demonstrates a snippet of a persuasive dialogue in which we are interested in. Under this setting, persuasion strategy recognition looks a bit similar to dialogue act recognition [5, 6, 7, 8], as dialogue acts and persuasion strategies both reflect speakers' intentions. However, they are different. Persuasion strategies are more complex than ordinary dialogue acts, and they are usually well-structured in dialogue and contain strong logic. To identify the persuasion strategies in a dialogue, we need a deeper understanding of conversation structures, semantic information of utterances, and even psychology attributes of speakers. Thus, persuasion strategy recognition in dialogues will be more challenging than dialogue act recognition.

Some existing models adopt LSTM-Softmax [9], hierarchical LSTM-CNN [10], and hybrid recurrent-CNN [11] to extract contextual features and predict labels. However, these methods completely depend on the hidden layers of the network which may lead to structural bias. Besides, speakers' responses will be influenced not only by the semantic history but also by the tactic history. Naturally, past strategies will influence future strategies. Although these methods have considered contextual correlations in the utterance level, they neglect the accompanied label dependencies in the tactic level. Moreover, intra-speaker dependencies have been neglected. Since the persuader's goal is clear in the persuasive dialogue, the persuader must organize his/her words strictly and logically during the persuasion process. As we can see in Table 1, the persuader carries out two consecutive credibility appeals by two utterances. If we don't look at the previous utterance from the persuader, we can hardly infer which strategy the latter utterance belongs to, as it merely looks like an answer to the persuadee's question. Therefore, intra-speaker features or self-dependencies can aid the model with

the understanding of logic inertia of individual speakers.

In this paper, we propose a transformer-based model coupled with Conditional Random Field (CRF) to model contextual understanding, inter-speaker, and label dependencies. On the benchmark dataset [11], our proposed approach surpasses strong baselines and state of the art showing efficacy of it over the existing approaches.

The paper is organized as follows: Section 2 discusses the related work on persuasion mining and user intent recognition; Section 3 elaborates the proposed framework; Section 4 illustrates the experiments; Section 5 shows the results and interprets the analysis, and finally Section 6 concludes the paper.

Role	Utterance	Annotation
ER	Do you ever donate to charity?	task-related-inquiry
EE	Yes, I support a few causes that I personally believe in very much.	positive-to-inquiry
ER	Have you ever heard of Save the Children?	source-related-inquiry
EE	Yes, but I don't know a lot about them.	positive-to-inquiry
EE	What is their mission?	ask-org-info
ER	Their mission is to promote children's rights, and provide relief and support to children in developing countries.	credibility-appeal
EE	That sounds interesting.	acknowledgement
EE	What countries do they work in?	ask-org-info
ER	They work in many countries across the world.	credibility-appeal
ER	For example, millions of children in Syria grow up facing the daily threat of violence.	emotion-appeal
ER	A donation could help these children greatly.	logical-appeal
EE	It sounds like it.	acknowledgement
EE	Do you donate to this charity?	ask-persuader-donation-intention
ER	I do.	self-modeling
ER	It is a great charity that does a lot of great work around the world.	logical-appeal
EE	Some charities are run better than others.	other

Table 1: A snippet of a persuasive dialogue where the annotations include persuasion strategies and non-strategy dialogue acts. ER and EE refer to the persuader and the persuadee respectively.

2. Related Work

Persuasive communication has been widely explored in various fields such as social psychology, advertising, and political campaigning. To get a better understanding of persuasiveness of requests on crowdfunding platforms, Yang et al. [12] presented a hierarchical neural network in a semi-supervised fashion to make the persuasiveness quantifiable. Egawa et al. [13] demonstrated five types of elementary units and two types of relations to characterize persuasive arguments and proposed an annotation scheme to capture the semantic roles of arguments in an online persuasive forum [14, 3, 4].

Furthermore, Hidey and McKeown [15] proposed a neural model with words, discourse relations, and semantic frames to predict persuasiveness in social media. Such previous work mainly focuses on evaluating persuasiveness in online forums, neglecting the psychological attributes of different speakers. Hence, in this work, we try to investigate persuasiveness in a conversation setting where persuasion goals, roles of persuader and persuadee as well as interactions between speakers are clearer.

Recent research on user intent recognition has shown promising results. For dialogue act (DA) recognition and classification, Khanpour et al. [9] presented a deep LSTM structure to classify dialogue acts in open-domain conversations. Liu et al. [10] incorporated contextual information for DA classification via a hierarchical deep learning framework. Also, Chen et al. [8] proposed a CRF-Attentive Structured Network where they captured hierarchical rich utterance representations to help improve DA recognition. For emotion recognition, DialogueRNN [16] and DialogueGCN [17] presented an RNN-based architecture and a GCN-based architecture to grasp hierarchical emotional information and speaker-level dependency. In our task, we try to recognize persuasive strategies utilized in a persuasive dialogue, where not only interactions between speakers make a difference to persuasive strategies but also whether the persuasion succeeds or not has an effect.

3. Methodology

3.1. Problem Definition

Given two interlocutors persuader and persuadee in a persuasion-driven dialogue $D = (u_1, ..., u_T)$ with T utterances, where utterance $u_t = (w_{t,1}, ..., w_{t,N_t})$ consists of a sequence of N_t words, the goal is to predict the persuasion strategy employed at each utterance. There are 10 different persuasion strategy categories for the persuader: 'logical appeal', 'emotion appeal', 'credibility appeal', 'foot-in-the-door', 'self-modeling', 'personal story', 'donation information', 'source-related inquiry', 'task-related inquiry' and 'personal-related inquiry'. And there are 12 persuasion response strategies for the persuadee: 'ask org info', 'ask donation procedure', 'positive reaction', 'neural reaction', 'negative reaction', 'agree donation', 'disagree donation', 'provide donation

amount', 'ask persuader donation intention', 'disagree donation more', 'task-related inquiry' and 'personal-related inquiry'. Except for these strategies, there is another category — 'non-strategy dialogue acts' for both persuader and persuadee. For convenience, here we also call the persuasion response strategy from the persuadee as the persuasion strategy.

3.2. Feature Extraction

We employ the RoBERTa model [18] to extract context-independent utterance level feature vectors. RoBERTa is a robustly optimized BERT [19] pretraining approach that uses two objective tasks: masked language modeling and next sentence prediction. RoBERTa uses the same network configuration as BERT which is based upon the widely used transformer architecture [20]. Several modifications from the BERT pretraining approach is proposed in RoBERTa, which leads to improvement in the end task performance. In particular, there are four key differences in the RoBERTa pretraining approach, which are: i) using dynamic masking instead of static masking, ii) using full sentences without next sentence prediction loss in the next sentence prediction task, iii) using larger mini-batch sizes during training, and iv) using a larger Byte-Pair Encoding (BPE) vocabulary size for tokenization. This modified pretraining procedure results in substantially improved performance in different auxiliary end tasks (GLUE, RACE, and SQuAD).

We fine-tune the RoBERTa Large model for persuasion strategy classification prediction from the transcript of the utterances. RoBERTa Large follows the original BERT Large architecture having 24 layers, 16 self-attention heads in each block, and a hidden dimension of 1024 resulting in a total of 355M parameters. Let an utterance u_t consists of a sequence of BPE tokenized tokens $w_{t,1}, w_{t,2}, ..., w_{t,N_t}$ and its strategy label is L_t. In this setting, the fine-tuning of the pretrained RoBERTa model is realized through a sentence classification task. A special token $[CLS]$ is appended at the beginning of the utterance to create the input sequence for the model: $[CLS], w_{t,1}, w_{t,2}, ..., w_{t,N_t}$. This sequence is passed through the model, and the activation from the last layer corresponding to the $[CLS]$ token is then used in a small feedforward network to classify it into its strategy label L_t.

Once, the model is fine-tuned for persuasion strategy classification, we pass the $[CLS]$ appended BPE tokenized utterances to the RoBERTa Large model and extract out activations from the final four layers corresponding to the $[CLS]$ token. These four vectors are then averaged to obtain the context-independent utterance feature vector having a dimension of 1024.

3.3. Our Model

Our model consists of three components: inter-speaker context encoder, speaker-specific context encoder, and strategy classifier. The encoders are based on transformers [20] and we employ conditional random field (CRF) [21] to classify the persuasion strategies. Fig. 1 shows the architecture of our framework.

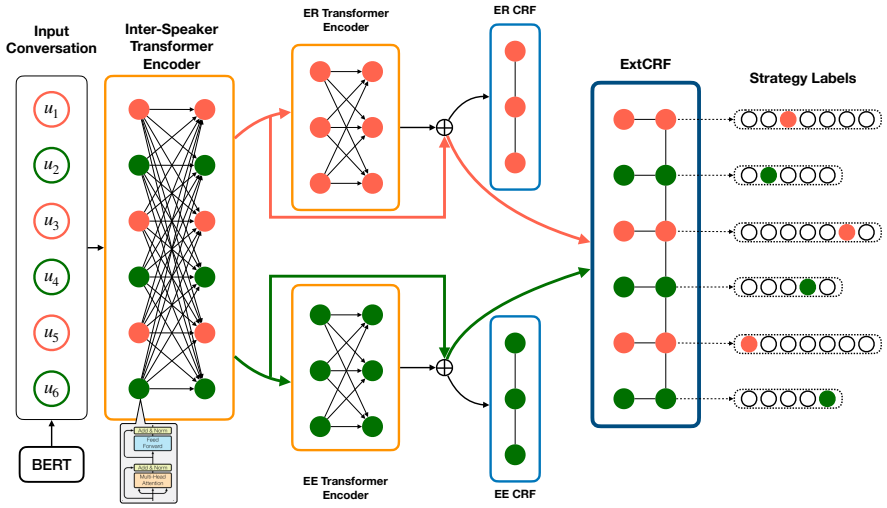

Figure 1: Architecture of our framework. ER and EE represent persuader and persuadee respectively, u represents utterance and $\oplus$ represents concatenation operation.

3.3.1. Inter-Speaker Context Encoder

Persuasive conversations flow along with the responses of persuader and persuadee. This sequence contains rich contextual information that can help us better understand the conversation. We feed the whole conversation to a transformer encoder to capture this inter-speaker contextual information.

As we illustrated in Section 3.2, we already obtained the context-independent utterance feature vectors. And the updated utterance representations in each dialogue are composed of these feature vectors. First, these representations $D' = (u'_1, u'_2, ..., u'_T)$ are mapped to queries Q, keys K and values V by linear projections with different weights:

$$\begin{aligned} Q_I &= W_{q_1} D' \\ K_I &= W_{k_1} D' \\ V_I &= W_{v_1} D' \end{aligned} \tag{1}$$

Then, we compute the dot products of the query with all keys to obtain the attention weight, and sum up all the weighted values to produce the context-aware output $Z \in \mathbb{R}^{T \times d_A}$:

$$\begin{aligned} Z_I &= Attention(Q_I, K_I, V_I) \\ &= \text{softmax}(\frac{Q_I K_I^{\mathrm{T}}}{\sqrt{d_{k_1}}}) V_I \end{aligned} \tag{2}$$

where d_{k_1} is the dimension of keys.

Next, the output Z is fed to a feedforward network which consists of a ReLU activation function and a linear activation function:

$$\begin{aligned} C_I &= FFN(Z_I) \\ &= max(0, Z_I W_{f_1} + b_{f_1}) W_{f_2} + b_{f_2} \end{aligned} \tag{3}$$

where W_* and b_* is the corresponding weight and bias respectively.

3.3.2. Speaker-Specific Context Encoder

Since, each interlocutor has his/her utterance logic. In this section, we model speaker-specific contextual information. In Section 3.3.1, we obtain a new sequential representation and in this section, we separate this sequence into two speaker-specific parts. Here, we define two notations — 0 represents the persuader and 1 represents the persuadee. Thus, the separated sequences can be written as $C_{I,0} = (u_{0,1}, ..., u_{0,T_0})$ and $C_{I,1} = (u_{1,1}, ..., u_{1,T_1})$.

Next, like what we have done in Section 3.3.1, we feed these two speaker-specific

sequences to a transformer encoder:

$$C'_{I,0} = TrsEncoder(C_{I,0}) \tag{4}$$

$$C'_{I,1} = TrsEncoder(C_{I,1}) \tag{5}$$

where the computing way of $TrsEncoder$ is the same as Eqs. (1) to (3).

3.3.3. Strategy Classification

We formulate this persuasion strategy classification as a sequence labeling problem. To capture the dependencies among strategy labels, we extend a linear-chain CRF (ExtCRF) to look at correlations between labels within neighborhoods in the inter-speaker sequence and do classification.

As we obtain two speaker-specific representations U'_0 and U'_1 from the speaker-specific encoders, we first concatenate them with the corresponding speaker-specific representations from the inter-speaker transformer encoder, and next merge these two sequences to one sequence:

$$C_{M,0} = C'_{I,0} \oplus C_{I,0} \tag{6}$$

$$C_{M,1} = C'_{I,1} \oplus C_{I,1} \tag{7}$$

$$C_M = merge(C_{M,0}, C_{M,1}) \tag{8}$$

where $\oplus$ is the concatenate operation and the $merge(*)$ operation merges two speaker-specific sequences $C_{M,0} = (c_{0,1}, ..., c_{0,T_0})$ and $C_{M,1} = (c_{1,1}, ..., c_{1,T_1})$ to one sequence $C_M = (c_1, ..., c_T)$ where $T = T_0 + T_1$ and the utterance representations come back to their original positions in the conversation.

ExtCRF classifier. Next, we feed the merged sequence C_M to our ExtCRF to classify the strategies. Formally, given a sequence of utterances $C_M = (c_1, ..., c_T)$, and the corresponding strategy sequence $Y_M = (y_1, ..., y_T)$, the probability of predicting the sequence of strategies can be written as:

$$P(Y_M|C_M) = \frac{1}{\mathbf{Z}(C_M)} \prod_{j=1}^{T} \phi_1(y_{j-1}, y_j)\phi_2(y_j, c_j) \tag{9}$$

$$\mathbf{Z}(C_M) = \sum_{y' \epsilon \mathcal{Y}} \prod_{j=1}^{T} \phi_1(y'_{j-1}, y'_j)\phi_2(y'_j, c_j) \tag{10}$$

where $\phi_1(*)$ and $\phi_2(*)$ are feature functions of the state transition potential and the emission potential, respectively. The state transition matrix provides us with the transition scores from label y_{j-1} to label y_j and it remains the same for each pair of consecutive time steps. The emission matrix provides us with the scores of label y_j at the j-th position of the strategy sequence.

$$\phi_1(y_{j-1}, y_j) = \exp(W^t_{y_{j-1}, y_j}) \tag{11}$$

$$\phi_2(y_j, c_j) = \exp(W^e_{y_j} c_j + b^e) \tag{12}$$

where $W^t_{y_{j-1}, y_j}$ provides the transition score from y_{j-1} to y_j and $W^e_{y_j}$ maps the context representation c_j to the feature score of y_j. Different from regular CRF, ExtCRF can deal with multiple label sets of various sizes. In the merged sequence, there are two different types of utterances, one uttered by the persuader and the other uttered by the persuadee. Thus, there exist four state transition cases: ER → ER, ER → EE, EE → ER and EE → EE. Accordingly, there are four types of transition matrices where the sizes are 11×11, 11×13, 13×11, and 13×13. 11 and 13 are the total number of labels for the persuader and the persuadee respectively. In our implementation, we integrated these four types of transition matrices into one 4D matrix which contains tag types, and each tag type records a transition matrix.

ER and EE CRF layers. Except for ExtCRF, here we also adopt a CRF layer to classify the strategies in speaker-specific sequences. Note that in our proposed model, we only take the results of ExtCRF to be the strategy predictions. For these two CRF layers in speaker-specific sequence, we merely add its cross-entropy to the objective function during training. Here the given sequences of utterances are $C_{M,0} = (c_{0,1}, ..., c_{0,T_0})$ for the persuader and $C_{M,1} = (c_{1,1}, ..., c_{1,T_1})$ for the persuadee, and the corresponding sequences of predicted labels are $Y_{M,0} = (y_{0,1}, ..., y_{0,T_0})$ and $Y_{M,1} = (y_{1,1}, ..., y_{1,T_1})$. Referring to Eqs. (9) to (12), we can obtain the probability of predicting the sequence of strategies. There is only one state transition within each CRF layer, and the transition matrices are of size 11×11 for ER CRF and 13×13 for EE CRF.

Persuasion result classification. Further, there is another auxiliary classifier in our completed framework. This classifier aims to predict whether the persuasion succeeds

or not. Here we first adopt self-attention to process the dialogue sequence and then apply a two-layer perceptron with a final softmax layer to predict the result:

$$l_t = \text{ReLU}(W_l D_t + b_l) \tag{13}$$

$$\mathcal{P}_t = \text{softmax}(W_{smax} l_t + b_{smax}) \tag{14}$$

$$\hat{y_t} = \underset{i}{\text{argmax}}(\mathcal{P}_t[i]) \tag{15}$$

where $\hat{y_t}$ is the predicted label for dialogue D_t. The cross-entropy of this classifier will be added to the objective function during training.

3.3.4. Model Training

We use the sum of cross-entropy from ExtCRF($\mathcal{L}_m$), ER CRF($\mathcal{L}_r$), EE CRF($\mathcal{L}_e$) and persuasion result classifier($\mathcal{L}_{succ}$) along with L2-regularization as the measure of loss($\mathcal{L}$), and our goal is to minimize the objective function during training:

$$\mathcal{L} = \mathcal{L}_m + \mathcal{L}_r + \mathcal{L}_e + \mathcal{L}_{succ} + \lambda \left\| \theta \right\|_2 \tag{16}$$

$$\mathcal{L}_{m,r,e} = -\frac{1}{\sum_{s=1}^{N} c(s)} \sum_{i=1}^{N} \sum_{j=1}^{c(i)} \log(P_{i,j}^{m,r,e}[y_{i,j}^{m,r,e}]) \tag{17}$$

$$\mathcal{L}_{succ} = -\frac{1}{\sum_{s=1}^{N} c(s)} \sum_{i=1}^{N} \log(P_{i}^{succ}[y_{i}^{succ}]) \tag{18}$$

where Eq. (17) illustrates the computing way of $\mathcal{L}_m$, $\mathcal{L}_r$ and $\mathcal{L}_e$, N is the number of samples/dialogues, $c(i)$ is the number of utterances in sample i, $P_{i,j}^{(*)}$ is the probability distribution of predicted labels for utterance j of dialogue i, $y_{i,j}^{(*)}$ is the expected class label for utterance j of dialogue i, λ is the L2-regularizer weight, and θ is the set of all trainable parameters within neural networks.

Additionally, at the time of testing in CRF layers, we adopt Viterbi algorithm [22] to obtain the optimal predicted sequence:

$$Y^* = \underset{Y}{\text{argmax}}(Y|C, \theta) \tag{19}$$

where Y is the sequence of predicted labels, C is the sequence of the given sequence of utterances, and θ is the set of all trainable parameters.

4. Experimental Setting

4.1. Dataset

The persuasive dialogue dataset used in our experiment is PERSUASIONFOR-GOOD [11] where one interlocutor aims to persuade the other interlocutor to donate his/her earning using different persuasion strategies. It consists of 1017 dialogues, where 300 dialogues are annotated with persuasion strategies. Specifically, there are average 10.43 turns per dialogue and on average 19.36 words per utterance. Also, this dataset provides actual donation made by the participants after the session ended. In this paper, we use these annotated dialogues to conduct our experiments and partition them into train and test sets with roughly 80/20 ratio.

4.2. Baselines

To obtain a comprehensive evaluation, we compare our model with the following baseline methods:

Hybrid RCNN [11]. This is the baseline along with PERSUASIONFORGOOD dataset. Hybrid RCNN extracts various features such as sentence embedding, context embedding, and sentence-level features via recurrent convolutional neural networks (RCNN) in this task. Experiments show that features from various aspects finally help improve the strategy classification.

Transformers [20]. Pre-training models have shown promising results on various NLP tasks. In our task, we provide a new baseline using transformers. First, we utilize BERT to extract independent utterance level feature vectors and employ a transformer encoder to encode the inter-speaker context and the speaker-specific context. Finally, we apply a two-layer perceptron with a final softmax layer to predict the strategies. The loss function contains losses from persuader and persuadee strategy classifiers and persuasion result classifier.

Transformers+CRF [20, 21]. This is a variant of the transformer model. A linear chain CRF layer is applied to speaker-specific context features to model the label dependencies and gives the optimal predicted sequence.

Models	ER	EE
	Macro F1	Macro F1
Hybrid RCNN	60.2	49.5
Transformers	64.1	50.9
Transformers+CRF	64.5	51.4
Transformers+ExtCRF	**66.4**	**52.2**

Table 2: Comparison with the baseline methods on PERSUASIONFORGOOD dataset. **ER** and **EE** represents predictions of persuader strategies and persuadee strategies respectively.

5. Results and Analysis

Utterance	Gold label	Pred. of our model	Pred. of transformers
ER:By directly asking for aid.	neutral-to-inquiry (Non)	Non	logical-appeal
EE:Thank you for your time.	thank(Non)	Non	disagree-donation
EE:What kind of children's charities do you know about?	task-related-inquiry	task-related-inquiry	ask-org-info
ER:Some of the causes they support include Emergencies (38%), Health and Nutrition (36%), and Education to more than 136 thousand children all over the world.	credibility-appeal	credibility-appeal	logical-appeal
ER:I am supposed to ask you if you care about people being killed in Syria and things like that, I don't want to cause you any emotional discomfort by talking about suffering people.	emotional-appeal	emotional-appeal	logical-appeal
EE:I would like to donate $0 but its not because I don't believe in the cause.	disagree-donation	disagree-donation	negative-reaction

Table 3: Samples in case studies. 'Non' represents non-strategy dialogue acts.

5.1. Comparison with the State of the Art

We compare our model with baseline methods for persuasion strategy classification in Table 2. As the dataset is highly imbalanced, here we only choose macro F1 to be the evaluation metric. We conducted five-fold cross-validation and used the average scores as the results. And we set the learning rate to be 0.0001 and L2 regularization weight to be 0.00001. Moreover, we utilized the validation set to tune the hyperparameters. Our model achieves new state-of-art macro F1 scores of 66.4% for the persuader strategy classification and 52.2% for the persuadee strategy classification, which outperforms all the baseline methods on PERSUASIONFORGOOD dataset.

To explain the improvement of performance, we should first figure out the nature of these models. The hybrid RCNN model utilizes recurrent neural networks (RNN)

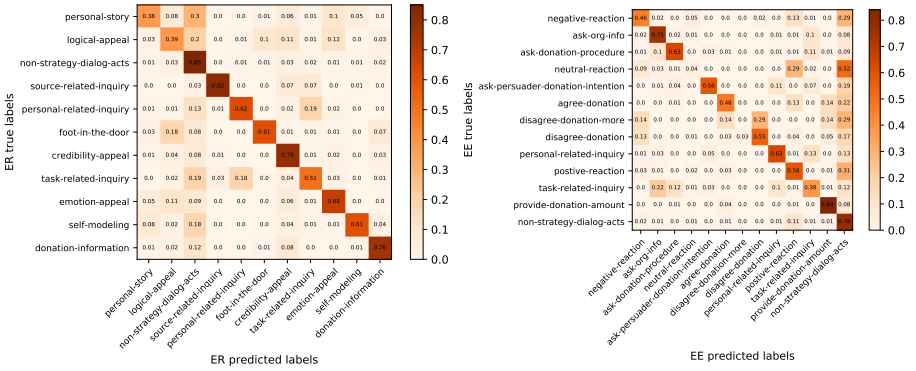

Figure 2: Confusion matrix of our model for (a) persuader strategy classification, and (b) persuadee strategy classification.

and convolutional neural networks (CNN) to extract the context features. Due to the limitations of RNN and CNN, they are not effective in encoding context information with complex semantics.

To encode those long sentences containing complex semantic meanings in persuasion conversations, we employ BERT to extract the features and feed them to transformer encoders. However, persuasion strategies are usually implicit. In Transformer+CRF, we add linear-chain CRF layers to the speaker-specific encoders, but the performance is similar to the one without CRF layers. In our proposed model, we first concatenate inter-speaker contextual features and intra-speaker contextual features and then extends a CRF layer to model the strategy dependencies in the merged sequence. The experiment shows an obvious improvement, with F1 scores increased by 1.9% and 0.8% respectively.

5.2. Case Studies

Specifically, we analyze the predictions of our model with ExtCRF and the state of the art. In Table 3, we list some cases comparing our method with the model of transformers. When encountering utterances that contain very little semantic information, e.g., non-strategy dialogue acts, our model maintains good performance while the transformer model does not. Moreover, we found our model performs better in the recognition of credibility appeal strategy combinations. There are several such strategy

combinations in the dataset and they usually appear after the 'ask org info' strategy from the persuadee. Generally in such combinations, the first one mainly replied to the persuadee and gave the information he/she asked, and the second one is what the persuader intended to express. In this case, semantic information alone is not enough. And as our proposed model considers the strategy transition, the predictions improve a lot. Further, there are some strategies like 'disagree-donation' and 'negative-reaction' that have something in common and are easy to be confused. In this case, the label dependencies have a positive effect on helping distinguish them.

However, we also observed some weaknesses in our model. In some cases, the persuadee was not willing to donate at first, but after persuasion, he/she agreed. Neither our model nor the transformer baseline has achieved satisfactory results. In these cases, there is usually a long distance between the 'disagree-donation' attitude and the 'agree-donation' attitude in a dialogue. Our model doesn't perform well in capturing long-distance strategy dependencies.

5.3. Ablation Study

Our ablation study includes two aspects, one for the classifiers in our proposed model and the other for our design of loss function. As shown in Table 2, after removing all the CRF components, we found that F1 scores of strategy classification for the persuader and the persuadee decreased from 66.4% and 52.2% to 64.1% and 50.9%, respectively. Moreover, if we substitute ExtCRF with regular CRF, we can still observe the F1 scores of the regular CRF model are 1.9% and 0.8% lower than those of our proposed model. This demonstrates our ExtCRF shows better performance on capturing strategy dependencies in sequences.

Further, we study the effect of our auxiliary losses in Table 4. First, we removed the loss of persuasion result classification and observed a decrease of 0.8% in the F1 score of persuader strategy classification. And when continuing to remove other losses from our loss function, we could also find a continuous decrease in the F1 score. This proves our losses have a positive effect on persuasion strategy classification.

Loss			ER
$\mathcal{L}_{ee}$	$\mathcal{L}_{merged}$	$\mathcal{L}_{succ}$	Macro F1
+	-	-	64.1
-	-	-	63.1
+	+	-	65.6
+	+	+	**66.4**

Table 4: Ablation results w.r.t losses that improve persuader strategy classification on PERSUASIONFORGOOD dataset.

5.4. Error Analysis

As shown in Fig. 2, we visualize the performance of our proposed model in two confusion matrices. In Section 5, we observed that 'personal story' tends to be misclassified into 'non-strategy dialog acts'. This is because utterances telling personal stories usually present an inconspicuous strategy tendency. Further, we found that several samples of 'logical appeal' are misclassified as 'emotional appeal' and 'credibility appeal'. One of the reasons is that one utterance may have multiple appeals. For instance, *'Save the Children is able to give away nearly everything they gather.'* This utterance can be classified into logical appeal since it tells the persuadee if he/she donates, the organization will probably help many young children. Also, it can be classified into credibility appeal since the organization tries to earn the persuadee's trust via this utterance.

Moreover, we observed there are more samples misclassified as 'non-strategy-dialog-acts' in persuadee strategy classification as shown in Section 5. For instance, the majority of samples of 'neutral-reaction' are misclassified as 'non-strategy-dialog-acts'. Similarly, one reason is that neutral reaction usually presents an inconspicuous strategy tendency. Further, we found samples of 'disagree-donation-more' are easily misclassified as 'disagree-donation'. We surmise this is due to the subtle difference between these two labels. Our model leaves some room for improvement to distinguish very similar labels.

6. Conclusion

In this paper, we have introduced a transformer-based neural network coupled with extended CRF, that captures both inter-speaker and intra-speaker contextual features and label dependencies to recognize persuasion strategies in dialogues. Through experiments on the benchmark dataset, we demonstrate the effectiveness of our proposed model in improving the performance of persuasion strategy recognition. Our analysis shows the strategy recognition task benefits from the label dependencies. Future work will focus on expanding the effectiveness of our proposed model to generate persuasive responses in dialogue systems.

References

[1] L. Ji, Z. Wei, X. Hu, Y. Liu, Q. Zhang, X.-J. Huang, Incorporating argument-level interactions for persuasion comments evaluation using co-attention model, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 3703–3714.

[2] T. Chakrabarty, C. Hidey, S. Muresan, K. Mckeown, A. Hwang, Ampersand: Argument mining for persuasive online discussions, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 2926–2936.

[3] C. Tan, V. Niculae, C. Danescu-Niculescu-Mizil, L. Lee, Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions, in: Proceedings of the 25th international conference on world wide web, International World Wide Web Conferences Steering Committee, 2016, pp. 613–624.

[4] C. Hidey, E. Musi, A. Hwang, S. Muresan, K. McKeown, Analyzing the semantic types of claims and premises in an online persuasive forum, in: Proceedings of the 4th Workshop on Argument Mining, 2017, pp. 11–21.

[5] L. Qin, W. Che, Y. Li, M. Ni, T. Liu, Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification, Thirty-Fourth AAAI Conference on Artificial Intelligence (2020).

[6] V. Raheja, J. Tetreault, Dialogue act classification with context-aware self-attention, arXiv preprint arXiv:1904.02594 (2019).

[7] T. Anikina, I. Kruijff-Korbayová, Dialogue act classification in team communication for robot assisted disaster response, in: Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, 2019, pp. 399–410.

[8] Z. Chen, R. Yang, Z. Zhao, D. Cai, X. He, Dialogue act recognition via crf-attentive structured network, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, ACM, 2018, pp. 225–234.

[9] H. Khanpour, N. Guntakandla, R. Nielsen, Dialogue act classification in domain-independent conversations using a deep recurrent neural network, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 2012–2021.

[10] Y. Liu, K. Han, Z. Tan, Y. Lei, Using context information for dialog act classification in dnn framework, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2170–2178.

[11] X. Wang, W. Shi, R. Kim, Y. Oh, S. Yang, J. Zhang, Z. Yu, Persuasion for good: Towards a personalized persuasive dialogue system for social good, arXiv preprint arXiv:1906.06725 (2019).

[12] D. Yang, J. Chen, Z. Yang, D. Jurafsky, E. Hovy, 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), 2019, pp. 3620–3630.

[13] R. Egawa, G. Morio, K. Fujita, Annotating and analyzing semantic role of elementary units and relations in online persuasive arguments, in: Proceedings

of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, 2019, pp. 422–428.

[14] Z. Wei, Y. Liu, Y. Li, Is this post persuasive? ranking argumentative comments in online forum, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016, pp. 195–200.

[15] C. Hidey, K. R. McKeown, Persuasive influence detection: The role of argument sequencing., in: AAAI, 2018, pp. 5173–5180.

[16] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, E. Cambria, Dialoguernn: An attentive rnn for emotion detection in conversations, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 6818–6825.

[17] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, A. Gelbukh, Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 154–164.

[18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[19] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

[21] J. D. Lafferty, A. McCallum, F. C. N. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proceedings of the

Eighteenth International Conference on Machine Learning, ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, p. 282–289.

[22] A. Viterbi, Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, IEEE transactions on Information Theory 13 (1967) 260–269.