DDRel: A New Dataset for Interpersonal Relation Classification in Dyadic Dialogues

Qi Jia, Hongru Huang, Kenny Q. Zhu[*]
Shanghai Jiao Tong University
Shanghai, China
Jia_qi@sjtu.edu.cn, onedesire@sjtu.edu.cn, kzhu@cs.sjtu.edu.cn

Abstract

Interpersonal language style shifting in dialogues is an interesting and almost instinctive ability of human. Understanding interpersonal relationship from language content is also a crucial step toward further understanding dialogues. Previous work mainly focuses on relation extraction between named entities in texts or within a single dialogue session. In this paper, we propose the task of relation classification of interlocutors based on their dialogues. We crawled movie scripts from IMSDb, and annotated the relation label for each session according to 13 pre-defined relationships. The annotated dataset DDRel consists of 6,300 dyadic dialogue sessions between 694 pairs of speakers with 53,126 utterances in total. We also construct session-level and pair-level relation classification tasks with widely-accepted baselines. The experimental results show that both tasks are challenging for existing models and the dataset will be useful for future research.

1 Introduction

Interpersonal relationship is an implicit but important feature underlying all dialogues, shaping how language is used and perceived during communication. People start to practice such style shifting in communication at very early stage unconsciously. Study (Dunn 2000) finds that when children listen to another person, their understanding of the counterpart depends on the nature of their relationship with the speaker. Study (Grabam, Barbato, and Perse 1993) finds that conversations between different partners are executed under different interpersonal motives, and thus the dialogues differ in topics and styles. Also, similar expressions may reflect different emotions and attitudes in different relationships.

Analyzing the relationship based on dialogues between interlocutors is well-motivated. First, it can provide dialogue systems with supplementary features for generating more suitable responses for different relationships, which helps in developing more intelligent role-playing chatbots. Second, it is useful in recommendation systems if the system can figure out the relationship between users according to their privacy-insensitive chats. Third, an automatic relationship classifier can help understand where the interpersonal stress/stimuli comes from in mental disorder treatment (Ariga et al. 2007; Karam et al. 2014; Yin et al. 2008). Besides, it can be also used for crime investigation, relieving the burden of manual monitoring and improving the productivity of searching in large amount of dialogue data.

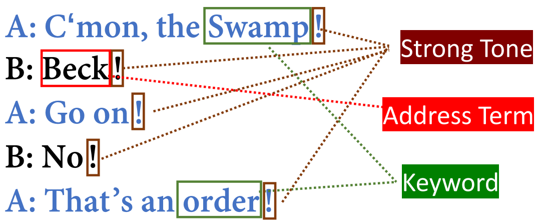


Figure 1: A sample dialogue and parts that are reported as informative by human testers.

Relation classification of dialogue sessions is not an easy task. Figure 1 shows a 5-turn dialogue example. We can see that it's challenging to fully contextualize such a short conversation without any prior knowledge, except one might infer that the two speakers are fellow soldiers in the military. Facing such problems, human usually resort to their communication experiences and commonsense knowledge to make sense of background stories and give inferences about the speakers' relationship. This shows that the inference of relationships from dialogues is possible but not straightforward for statistical models.

Previous work mainly focuses on relation classification between named entities. Sentence-level relation classification datasets, such as FewRel (Han et al. 2018) and TACRED (Zhang et al. 2017), have been widely studied (Zhang, Qi, and Manning 2018; Gao et al. 2019; Zhang et al. 2019), targeting on figuring out the correct relation type between entities within a given sentence. Recent research sets sights on inter-sentence relations. DocRED (Yao et al. 2019) has been proposed as the largest document-level relation extraction dataset from plain text, where person (an entity type) only occupies 18.5% of the entities. DialogRE (Yu et al. 2020) aims at predicting the relations between two arguments and MPDD (Chen, Huang, and Chen 2020) is a Chinese dataset for predicting relations between the speaker and listeners of each turn in a dialogue session. However,

[*] Corresponding author.

relations in both datasets are limited to the current dialogue session, without cross-session considerations.

Different from their problem definition, we are trying to figure out the interpersonal relationships between speakers in dyadic dialogues from two perspectives, session-level and pair-level. Multiple dialogue sessions may happen between each pair of speakers and it is usually not easy to figure out the relationship between a pair of speakers with only one session and no background context. Human has the ability to make the connections between multiple sessions and construct the whole picture between two speakers. In other words, cross-session inferences are required for predictions.

In this paper, we propose a new dyadic dialogue dataset for interpersonal relation classification called DDRel. The dataset consists of 6300 dialogue sessions from movie scripts crawled from IMSDb between 694 pairs of speakers, annotated with relationship labels by human. 13 relation types according to Reis and Sprecher (Reis and Sprecher 2009) are covered in our dataset and these types can cover most of the interpersonal relations in our daily life. Several strong baselines and human evaluations are implemented. The results and future work of our dataset are discussed.

In summary, this paper makes following contributions:

- We propose the task of dialogue relation classification for speakers, different from the previous intra-sentence or inter-sentence relation classification tasks (Sec. 2).
- To the best of our knowledge, we construct the first-ever dialogue relation classification dataset for analyzing interpersonal relationships between speakers with multiple dialogue sessions (Sec. 3).
- We establish a set of classification baselines on our dataset using standard learning-based techniques. The gap between SOTA models and human performances show the difficulty of this task and higher requirements for current models (Sec. 4 and Sec. 5).

2 Task Definition

Our work aims at identifying the interpersonal relation between interlocutors in dyadic dialogues. The types of relationships are pre-defined, annotated as $R = \{R_1, R_2, ..., R_m\}$ where m is the number of relation types. A number of sessions may happen between the same pair of interlocutors. So, we define the relation classification task in two levels: session-level and pair-level.

Given the j-th dialogue session D_j^i between the i-th pair of interlocutors, **session-level relation classification task** is to inference the most possible relation type for this session:

$$R_j^i = \arg\max_R f_s(D_j^i) \qquad (1)$$

Due to the fact that it's quite hard for even human to fabricate the whole story only through one dialogue session, **pair-level relation classification task** is defined as follows. Given dialogues between the i-th pair of interlocutors denoted as $D^i = (D_1^i, D_2^i, ..., D_n^i)$, pair-level relation classification task is to figure out the most possible relation type for this pair, i.e.:

$$R^i = \arg\max_R f_p(D^i) = \arg\max_R f_p(D_1^i, D_2^i, ..., D_n^i) \quad (2)$$

13-class taxonomy of relationships is covered in our DDRel dataset, including *child-parent*, *child-other family elder*, *siblings*, *spouse*, *lovers*, *courtship*, *friends*, *neighbors*, *roommates*, *workplace superior-subordinate*, *colleagues*, *opponents* and *professional contacts*, based on Reis and Sprecher (Reis and Sprecher 2009), in which they elaborate on psychological and social aspects of various relationships. We define these categories by social connections because they make general sense in life. Although individual difference exists in every real-world case, it was found that such relationship category has different expectations, special properties (e.g., marriage usually involves sex, shared assets and raising children) (Argyle and Furnham 1983), distinctive activities (e.g., talking, eating, drinking and joint leisure for friendship) and rules (Argyle and Henderson 1984) of its own, which are agreed across cultures. Note that this is not an all-round coverage of all possible relationships in human society and we aim to cover those common ones in real life which may be of interest in interpersonal relationship research. These fine-grained labels are prepared for possible related future research.

To evaluate the classification ability of the model from coarse-grained to fine-grained, we also cluster our 13 specific relation types into 6 classes and 4 classes considering the social field, the seniority and the closeness between two speakers. The details of relation types are listed in Table 1.

3 Dataset

Although there are many currently available dialogue datasets, most of them are used for training automatic dialogue robots/systems, thus they either do not cover the diversity of interpersonal relationships, or do not come with relationship labels. Therefore, we build a new dataset composed of $6,300$ sessions of dyadic dialogues with interpersonal relationship labels between two speakers, extracted from movie scripts crawled from the Internet Movie Script Database (IMSDb). More details are explained as follows.

3.1 Dataset Extraction and Processing

Initially, we crawl 995 movie scripts from IMSDb, and 941 of them remain after we automatically match the titles with movies in IMDb[1] and filter out those that do not meet following requirements: 1) Don't have a match in IMDb; 2) Not in English; 3) Very unpopular(measured by number of raters).

By observing the formats of the scripts and manually defining text patterns, we split each script into scenes, extract the sequence of (speaker, utterance) pairs for each scene and identify subsequences that meet the following requirements as dyadic dialogue sessions:

- Two speakers speak alternately without being interrupted by a third one;
- Each dialogue session contains at least 3 turns.

We set this minimum length requirement to make sure that two speakers are speaking to each other instead of participating in a group discussion. Finally, we count the total number

[1] https://www.imdb.com/

4 classes	6 classes	13 classes	# Sessions	% Sessions	# Pairs	% Pairs	# Turns	% Turns
Family	Elder-Junior	Child-Parent	414	6.57	67	9.65	3,377	6.36
		Child-Other Family Elder	91	1.44	12	1.73	632	1.19
	Peer	Siblings	211	3.35	27	3.89	1,585	2.98
		Spouse	568	9.02	51	7.34	4,784	9.01
Intimacy	Intimacy	Lovers	1,852	29.40	244	20.75	17,474	32.89
		Courtship	146	2.32	15	2.16	1,323	2.49
Others	Peer	Friends	1,049	16.65	124	17.87	8,900	16.75
		Neighbors	21	0.33	2	0.29	189	0.36
		Roommates	120	1.90	8	1.15	966	1.82
Official	Elder-Junior	Workplace Superior-Subordinate	536	8.51	79	11.38	3,958	7.45
	Peer	Colleague/Partners	710	11.27	76	10.95	5,455	10.27
		Opponents	203	3.22	33	4.76	1,532	2.88
		Professional Contact	56	8.07	56	8.07	2,952	5.56

Table 1: Statistics on categories of interpersonal relation types.

of turns taken between each pair and filter out those having fewer than 20 turns to make sure the relationship between the two speakers is significant and not as trivial as greetings between strangers. This filtering step also helps reduce the cost of labeling because more sessions can share the same pair of speakers.

3.2 Annotation Procedure

Although interpersonal relationships are not static or mutual exclusive, most of them exhibit relative stability over time (Gadde and Mattsson 1987), and relationships in movies are usually more clear-cut. Therefore, in this paper, we model relationship as a single stable label. Such assumption simplifies our task and significantly reduces the workload of labeling, though it introduces ambiguity in certain cases such as evolving relationships (e.g., courtship → lover → spouse) or concurrent ones that do not usually exist together(e.g., enemies falling in love). To avoid these situations, we require the annotator to only assign labels when the relationship is clear, relatively stable and typical.

Our ground truth annotator was provided with the movie title, the pair of characters involved in the dialogue, movie synopsis from IMDb and Wikipedia for each movie, as well as complete access to the Internet, and was asked to choose between one out of tens of classes mentioned in Reis and Sprecher's work (Reis and Sprecher 2009) or "Not applicable (NA)" label. It took the annotator 100 hours across one and a half months to finish the annotation of 300 movies, at a rate of approximately 4.07 minutes per pair. Only 47.11% of the pairs received a specific label, while others are considered "not applicable". Finally, 13 kinds of relation types are labeled in our dataset, covering a variety of interpersonal relationships and enough for developing classification methods on this task.

Second-Annotator Verification Due to excessive costs of the annotation task, we are not able to commit multiple annotators. To compensate that, we verify the accuracy of annotation by having a second person label 100 pairs with the same experimental settings. The inter-annotator agreement (kappa) is 82.3% for 13-classes. This indicates that incorrect labels are limited, and the annotation by the first human is reliable.

3.3 Dataset Statistics

The current version of the DDRel dataset [2] contains $6,300$ labeled sessions of dyadic dialogues, taking place between 694 pairs of interlocutors across 300 movies. The average number of turns in each dialogue is 8.43, while it varies greatly (the standard deviation is 6.94). The number of sessions for each pair of interlocutors also varies a lot with $avg = 9.08$ and $std = 7.80$. The whole dataset is split into train/development/test sets by 8:1:1 as shown in Table 2. All of the dialogue sessions between the same interlocutors are assigned to the same subset and there is no overlap between three subsets.

The distribution of the whole dataset on 13 relation types are shown in Table 1. *Lovers*, *Friends* and *Colleague/Partners* are the three largest classes and take up about half of the dataset, while the smallest relation type *Neighbor* only has 2 pairs of interlocutors with 21 dialogue sessions. The proportion of different relation types are unbalanced, aggravating the difficulty of classification tasks.

	train	development	test
# Pair of Speakers	541	75	78
# Sessions	5,037	653	610
# Turns	42,564	5,210	5,352
Sessions per pair (mean)	9.31	8.71	7.82
Sessions per pair (std)	8.18	6.35	5.96
Turns per session (mean)	8.45	7.98	8.77
Turns per session (std)	6.96	5.60	7.93

Table 2: Statistics on the splitted datasets.

4 Experiments

In this section, we introduce the baseline models, human evaluation settings and evaluation metrics.

[2] We have processed 941 scripts and manually labeled 300 of them with relationships at present.

4.1 Baseline Models

We introduce two naive baseline methods, Random and Majority, and three strong neural baseline models, CNN, LSTM and BERT. The code and dataset are avaliable at Github [3].

Random: A relation type is randomly assigned to a dialogue session or a pair of interlocutors.

Majority: The most frequent relation type is assigned to a dialogue session or a pair of interlocutors.

CNN: TextCNN, proposed by Kim (2014), is a strong text classification model based on convolution neural network. All of the utterances in a dialogue session are concatenated as the input to the embedding layer, where 300-dimension pre-trained Glove (2014) embeddings are used and freezed during training. Following the setting of Kim (2014), we use three convolution layers with kernel size equaling 3, 4, and 5 to extract semantic information from the dialogue. A dropout layer with probability 0.5 is attached to each convolutional layer to prevent overfitting. Finally, a linear layer and a softmax function are set for the final prediction. The loss function is the negative log likelihood loss. Stochastic gradient descent is used for parameter optimization with the learning rate equaling 0.01.

LSTM: The attention-based bidirectional LSTM network by Zhou et al. (2016) is implemented as another neural baseline. The same pre-trained Glove embeddings are used for the embedding layer. Then high-level features are extracted by a single Bi-LSTM layer. The last hidden states of both directions are concatenated as the query to do the self-attention among the input words. Finally, the weighted summed feature vector can be used to characterize the whole session and used for the final relation classification with a linear layer and a softmax function. We use AdamDelta as optimizer with learning rate 0.0003.

BERT: We fine-tune the base model of BERT released by Devlin et al. (2019). All of the utterances in a dialogue session are also concatenated with the special token [CLS] at the start of the sequence. Following the general procedure of fine-tuning BERT, we pass the output hidden state of [CLS] token into a fully-connected layer for classification and use Adam as optimizer with learning rate $1e-6$. We fine-tune the model for 32 epochs at most with early stopping patience equaling 3.

The above baselines can be directly used for session-level classifications. For pair-level classifications, we do the following calculation for each neural baseline based on the session-level trained models: We first calculate the MRR metric of each relation type for each session. Then, given a pair of interlocutors with multiple sessions, the confidence score for each relation type can be regarded as the average MRR among sessions. The relation type with the maximum confidence score is the final prediction for this pair.

4.2 Human Evaluation Settings

To give an upper bound of our proposed DDRel dataset, we hired human annotators to do the relation classification tasks on the test set. Since given the 13-class classification results, the 4-class or 6-class classification results are obvious for human. We only asked annotators to do the 13-class interpersonal relation classification tasks.

We asked 2 volunteers to do the 13-class relationship task on session-level samples. Each session is showed individually and volunteers are required to choose the most possible relation type. Another 2 volunteers are hired to do the 13-class relationship task on pair-level samples. All of the dialogue sessions between a pair of speakers are given to the volunteers to inference the relation types.

4.3 Evaluation Metrics

Each classification model is trained separately for relation classification tasks on different granularity. We use accuracy and F1-macro scores for evaluation.

5 Results and Discussions

In this section, we discuss the classification performances and provide a simple data augmentation method for pair-level classification, with a case study and future directions.

5.1 Session-level Performance

We run all of the baseline models with three different random seeds and then obtain the mean and std values of the evaluation metrics. The results of baselines and human upper bound on session-level relation classification tasks are shown in Table 3. The difficulties on session-level tasks are in proportional to the number of classes, as the performances of all of the models and human annotators decrease from 4-class task to 13-class task. The decreases are about 10% and 20% on accuracy for models and human respectively.

For Majority, the accuracy is even higher than a neural baseline (LSTM), while the F1-macro is the lowest due to the unbalanced data distribution between classes. The neural baselines mostly perform better than Random and Majority.

The comparison of performances on neural baselines is BERT>CNN>LSTM. LSTM is much weaker than CNN. The gap between them on 13-class classification is smaller than 4-class and 6-class classifications due to the fact that both of them failed on fine-grained classification tasks. The F1-macro is only 4.63% and 9.20% respectively with high variance. BERT, a pre-trained language model baseline, performs much better and more stable than the other two neural models with higher scores and lower variance. The gaps between evaluation metrics are also smaller than other baselines, indicating that it can handle the problem of unbalanced data to some extent.

The human performance is the average score of two annotators. We calculate the Cohen's Kappa between them, and the agreement is 0.429, 0.336 and 0.301 for 4-class, 6-class and 13-class relation classification tasks respectively. The agreement on 6-class and 13-class tasks is fair, and on more coarse-grained task, 4-class task, is moderate. It' s also quite difficult for human to identify the relationship between interlocutors based on only one session. It seems that BERT outperforms human upper bound on accuracy of the 13-class classification task, but actually there is no significant difference between them, and F1-macro score of human upper bound is statistically significantly better than BERT with p-value less than 0.05.

[3] https://github.com/JiaQiSJTU/DialogueRelationClassification

		4-class		**6-class**		**13-class**	
		Acc	F1-macro	Acc	F1-macro	Acc	F1-macro
Session-level	Random	23.00±3.56	22.67±3.71	17.33±2.62	15.80±3.00	8.33±2.62	6.63±2.12
	Majority	31.00±0.00	11.80±0.00	31.00±0.00	7.90±0.00	26.00±0.00	3.20±0.00
	LSTM	29.80±1.28	22.87±1.24	30.83±1.16	11.10±0.08	28.50±1.44	4.63±0.45
	CNN	42.67±2.93	33.27±6.63	37.80±1.31	31.40±6.67	32.33±2.46	9.20±4.97
	BERT	47.10±1.28	44.53±1.10	41.87±0.81	39.40±0.85	39.40±0.36	20.40±0.67
	Human	56.00±6.00	55.20±6.30	50.00±9.00	53.00±8.10	38.50±5.50	40.75±8.15
Pair-level	Random	28.20±9.30	26.90±9.24	17.93±7.89	16.20±7.54	6.43±2.76	5.73±2.64
	Majority	23.10±0.00	9.40±0.00	23.10±0.00	6.20±0.00	19.20±0.00	2.50±0.00
	LSTM	25.63±2.76	13.13±5.06	22.67±0.61	6.40±0.29	19.20±0.00	2.57±0.05
	CNN	47.47±2.76	35.03±5.80	38.47±4.21	30.40±9.06	22.20±6.08	7.07±6.04
	BERT	58.13±0.61	52.00±0.86	42.33±2.76	38.00±1.14	39.73±1.79	24.07±0.63
	Human	75.65±3.85	73.00±4.40	72.40±4.50	73.55±5.45	63.45±1.95	54.40±3.00

Table 3: The classification results(%) on session-level tasks and pair-level tasks.

5.2 Pair-level Performance

Table 3 also includes the results on pair-level tasks. The difficulties between classification tasks on different granularity are the same as session-level classification tasks and the comparisons between baseline models are also the same: BERT>CNN>LSTM.

By using the MRR metric to get the prediction on pair-level performance of each model as explained in Section 4.1, it aggravates the polarization of the model performances. The strong baselines like CNN (except on 13-class classification task) and BERT achieve higher scores on pair-level tasks, while others, including LSTM and 13-class CNN model, perform even worse on pair-level tasks. The performance of LSTM is close to Majority baseline. The reason for this phenomenon is that we only assign one label for multiple sessions on pair-level tasks. If the model is weak, it tends to give some extremely unreasonable predictions on some of the sessions for a given pair of interlocutors. Even though there may be some correct predictions on session level, the final prediction for this pair is wrong. On the other hand, if the model is strong, it can give more reasonable predictions for most sessions. Then although there may be some wrong cases on session level, they will be tolerated. The performance of these models will increase.

The gap between CNN and BERT decreases from 4-class to 6-class while increases greatly from 6-class to 13-class. The convolution-based model seems more stable on coarse-grained tasks, and drops dramatically on the 13-class fine-grained task. On the contrary, the performance of the fine-tuned language model decreases rapidly from 4-class to 6-class and decreases slowly from 6-class to 13-class. As a result, the advantage of BERT model on 6-class classification tasks is limited beyond the CNN baseline.

Human annotators also performance much better on pair-level tasks than session-level tasks. The Cohen's Kappa for two annotators is 0.698, 0.687 and 0.614 for 4-class, 6-class and 13-class classification tasks respectively, showing substantial agreements. The higher performances and agreements are consistent with the intuition that, with multiple sessions for a given pair, human are able to find more correlations between sessions and better understand the background of two interlocutors. In this way, we think pair-level relation classification tasks are more reasonable, challenging and meaningful for the development of current models.

The gap between best baseline BERT and Human performances also shows the limitation of current models. We draw the confusion matrix for the best neural model BERT and human performances in Figure 2. We can see that for coarse-level relation classification tasks, the performances of BERT and human have some similarities. They both did well on predicting the official relation type on 4-level task, and intimacy-peer relation type and official-peer relation type on 6-level task. For 13-classification tasks, BERT fails dramatically which may due to the unbalanced data distribution, tending to predict the relation type of "*lovers*", while human perform well on "*Workplace Superior-Subordinate*".

5.3 A First Step on Cross-session Consideration

For pair-level classification tasks, our neural baselines give the final predictions by aggregating the predictions for each session. In this way, some interactions between sessions may be omitted. To clarify the existence of cross-session interaction in our DDRel dataset, we augment the original pair-level samples with multiple sessions as follows: i) Cut the session into K pieces according to its length (the number of utterances). ii) Concatenate the session pieces at the same cut point in two consecutive sessions to generate a new session for the given pair. It should be noted that the order of sessions in each pair follows the chronological order. For example, in Figure 3 there is a pair with two sessions. Each session is cut into 3 pieces when $K = 3$ and we get two augmented sessions for this pair.

Using the best BERT model we trained above, we augment the test set and re-evaluate the pair-level prediction results in Table 4. Most classification results are increased by augmentation operation. We further augment all of the datasets in DDRel, and train and test the BERT baseline on the augmented dataset with $K = 3$. The results on 6-class and 13-class enhanced significantly with accuracy equaling 49.13% and 41.87% respectively, and with F1-macro equaling 46.93% and 26.83% respectively. All of the results indi-

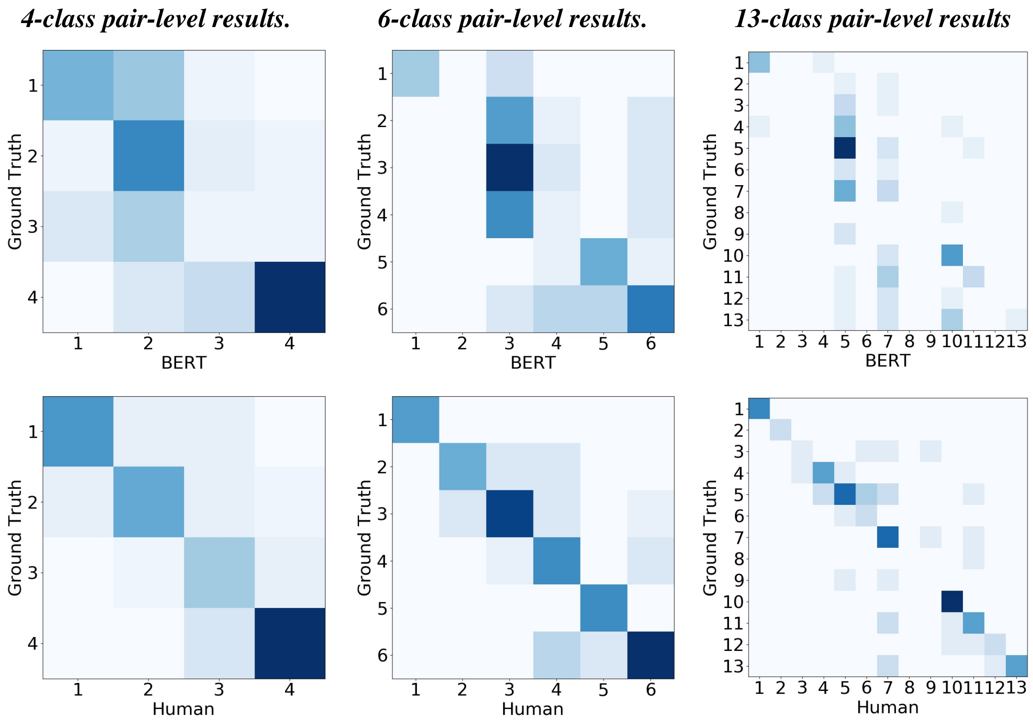


Figure 2: The confusion matrix of relation classification tasks.

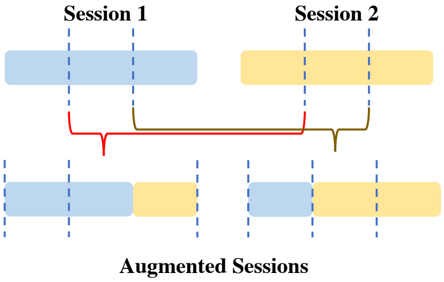


Figure 3: An illustration of data augmentation when $K = 3$.

	4-class		6-class		13-class	
	Acc	F1	Acc	F1	Acc	F1
$K = 2$	+1.73	+2.50	-0.40	-0.97	+2.13	+0.63
$K = 3$	+0.43	+1.33	+1.70	+0.20	+2.13	+0.80
$K = 4$	-0.43	+0.43	+2.56	+1.43	+2.13	+0.90

Table 4: The pair-level classification results(%) with data augmentation on the test set compared with BERT baseline. K is the hyper-parameter for our proposed data augmentation method. F1 is short for F1-macro.

cate the existence and importance of cross-session information for pair-level relationship classifications.

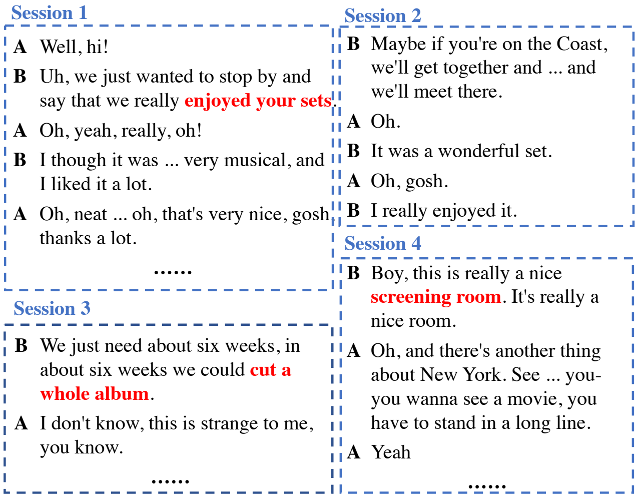
5.4 Case Study and Future Directions

We show a representative pair-level classification case with 4 sessions in Figure 4. As for human, we can make inferences according to keywords such as “enjoyed your sets”, “cut a whole album” and “screening room”. Based on our background experience and knowledge, such conversations are more likely happening between two guys with cooperation in making music. The cues here are not obvious in a single session but are very assuring when four sessions are considered together. Human choose “Official” while the best baseline BERT mistakes it to be “Intimacy” for this sample. The model may be confused by the informal expressions and emotional words such as “wonderful”.

According to the case study, we consider the further research on this task as follows:

Cross-session Consideration. As talked about in Section 5.2, classification of a pair of interlocutors based on multiple sessions between them is a more reasonable and meaningful task. Due to the fact that the number of sessions for pairs varies a lot and it’s difficult and unreasonable to concatenate all of the utterances in these sessions as the input for models, we only combines the prediction results of each session to get the final pair-level predictions and made a simple step on cross-session consideration by data augmentation. Developing models that could better find the cues between sessions is an important direction for current models.

Commonsense Knowledge. Another limitation of current models is due to the lack of commonsense knowledge, even for commonly pre-trained language models. Human can better inference the background of two interlocutors with the previous stories or experiences they have had. Further pre-training the language models on more similar cor-



Session 1

A Well, hi!

B Uh, we just wanted to stop by and say that we really **enjoyed your sets**.

A Oh, yeah, really, oh!

B I though it was ... very musical, and I liked it a lot.

A Oh, neat ... oh, that's very nice, gosh, thanks a lot.

......

Session 2

B Maybe if you're on the Coast, we'll get together and ... and we'll meet there.

A Oh.

B It was a wonderful set.

A Oh, gosh.

B I really enjoyed it.

Session 3

B We just need about six weeks, in about six weeks we could **cut a whole album**.

A I don't know, this is strange to me, you know.

......

Session 4

B Boy, this is really a nice **screening room**. It's really a nice room.

A Oh, and there's another thing about New York. See ... you-you wanna see a movie, you have to stand in a long line.

A Yeah

......

Figure 4: A pair-level case with 4 sessions. The words colored in red are possible classification cues.

pus and incorporating the commonsense knowledge, such as ConceptNet (Speer, Chin, and Havasi 2017), are possible solutions.

6 Related Work

Related work on relation classification and dialogue datasets is discussed in this section.

6.1 Relation Classification

Relation classification or extraction is an important first step for constructing structured knowledge graph in NLP with popular benchmark datasets such as NTY-10 (Riedel, Yao, and McCallum 2010) and the SemEval-2010 dataset (Hendrickx et al. 2010). Previous datasets for relation classification focus on figuring out the relation type between two entities in a single sentence, including FewRel (Han et al. 2018) and TACRED (Zhang et al. 2017). However, such intra-sentence relation classification has a limitation in real applications and looses nearly 40.7% of relational facts according to previous research (Swampillai and Stevenson 2010; Verga, Strubell, and McCallum 2018; Yao et al. 2019).

Inter-sentence relation classification or document-level relation classification has gained more attention in recent years. There are only several small-sized dataset for this task, including a specific-domain dataset PubMed (Li et al. 2016) and two distant supervised datasets from Quirk and Poon (2017) and Peng et al. (2017). To facilitate the research in this area, DocRED (Yao et al. 2019) has been proposed as the largest dataset for document-level relation classification. Our task is different from it since we focus on interpersonal relations while person-related entities is only a small component in DocRED. Besides, our task is based on dialogue sessions instead of plain documents and interpersonal relation classification may need inferences beyond session level.

6.2 Dialogue Datasets

Dialogue system is a hot research point in recent years with a rapid growing number of available dialogue datasets. Generally, dialogue datasets can be divided into two categories. One is the task-oriented dialogue datasets such as Movie Booking Dataset (Li et al. 2017a), CamRest676 (Ultes et al. 2017) and MultiWOZ (Budzianowski et al. 2018). These datasets focus on single or multiple targeting domains and are usually labeled with dialogue act information, serving for the slot filling (Liu et al. 2020) and dialogue management tasks (Budzianowski and Vulic 2019) when building task-oriented dialogue systems. The other is the open-domain chit-chat datasets such as DailyDialog (Li et al. 2017b), MELD (Poria et al. 2019) and PERSONA-CHAT (Zhang et al. 2018). The resource of these conversations is usually social media platforms, including Facebook, Twitter, Youtube, and Reddit. Researches on these datasets mainly focus on emotion recognition and emotion interplay among interlocutors, helping chatbots generate more emotionally coherent (Ghosal et al. 2019) and persona consistent responses (Zheng et al. 2020).

There are two existing datasets similar to our settings. One dataset is the DialogRE (Yao et al. 2019). It focuses on predicting the relations between two arguments in a dialogue session, where relations between arguments of interlocutors are rare. Also, since all of the 1,788 dialogue sessions are crawled from the transcript of *Friends*, it suffers a limitation of the diversity of scenarios and speakers. Another dataset is MPDD (Chen, Huang, and Chen 2020). This dataset contains 4,142 dialogues annotated with speaker-listener interpersonal relation in multi-party dialogues for each utterance, while the relation types in our dataset is not such directional relationships. Besides, both datasets ignore the fact that interlocutors may have multiple sessions which is considered in our task and dataset. Our task is more reasonable with practical social meanings.

There are also similarities between our task and dialogue summarization (Misra et al. 2015; Gliwa et al. 2019), where they are both required to pick up the useful information throughout the given texts. Our pair-level relation classification setting with multiple sessions have similarities with multi-document summarization (Lin and Hovy 2002). However, summarization usually refers to a natural language generation task, whereas our task is defined as a classification task. Both extractive and abstractive summarization models can not be directly used for this task, but future work may take advantage of this viewpoint.

7 Conclusion

This paper proposes the interpersonal relation classification task for interlocutors in dyadic dialogues, accompanied with a new reasonable sized dialogue dataset called DDRel. The cross-session relation classification is raised for the first time and the results of baseline models show the limitation for current methods on this new task. Models that taking advantages of multiple sessions and commonsense knowledge are expected to be explored as future work.

Acknowledgements

Sincere gratitude goes to Minxue (Sandy) Niu for her initial contribution and subsequent assistance in preparing this paper. This research was supported by the SJTU-CMBCC Joint Research Scheme, SJTU Medicine-Engineering Cross-disciplinary Research Scheme, and NSFC grant 91646205.

References

Argyle, M.; and Furnham, A. 1983. Sources of satisfaction and conflict in long-term relationships. *Journal of Marriage and the Family* 481–493.

Argyle, M.; and Henderson, M. 1984. The rules of friendship. *Journal of social and personal relationships* 1(2): 211–237.

Ariga, M.; Yano, Y.; Doki, S.; and Okuma, S. 2007. Mental Tension Detection in the Speech based on physiological monitoring. In *IEEE International Conference on Systems, Man and Cybernetics*, 2022–2027.

Budzianowski, P.; and Vulic, I. 2019. Hello, It's GPT-2 - How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems. In *Proceedings of the 3rd Workshop on Neural Generation and Translation@EMNLP-IJCNLP*, 15–22.

Budzianowski, P.; Wen, T.; Tseng, B.; Casanueva, I.; Ultes, S.; Ramadan, O.; and Gasic, M. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.

Chen, Y.; Huang, H.; and Chen, H. 2020. MPDD: A Multi-Party Dialogue Dataset for Analysis of Emotions and Interpersonal Relationships. 610–614. Proceedings of The 12th Language Resources and Evaluation Conference.

Devlin, J.; Chang, M.; Lee, K.; and 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 (Volume 1 Long and Short Papers)*, 4171–4186.

Dunn, J. 2000. Mind-reading, emotion understanding, and relationships. *International Journal of Behavioral Development* 24: 142–144.

Gadde, L.-E.; and Mattsson, L.-G. 1987. Stability and change in network relationships. *International journal of research in marketing* 4(1): 29–41.

Gao, T.; Han, X.; Liu, Z.; and Sun, M. 2019. Hybrid Attention-Based Prototypical Networks for Noisy Few-Shot Relation Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6407–6414.

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. F. 2019. 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*, 154–164.

Gliwa, B.; Mochol, I.; Biesek, M.; and Wawer, A. 2019. SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, 70–79.

Grabam, E. E.; Barbato, C. A.; and Perse, E. M. 1993. The Interpersonal Communication Motives Model. *Communication Quarterly* 41: 172–186.

Han, X.; Zhu, H.; Yu, P.; Wang, Z.; Yao, Y.; Liu, Z.; and Sun, M. 2018. FewRel: A Large-Scale Supervised Few-shot Relation Classification Dataset with State-of-the-Art Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4803–4809.

Hendrickx, I.; Kim, S. N.; Kozareva, Z.; Nakov, P.; Séaghdha, D. Ó.; Padó, S.; Pennacchiotti, M.; Romano, L.; and Szpakowicz, S. 2010. SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 33–38.

Karam, Z. N.; Provost, E. M.; Singh, S.; Montgomery, J.; Archer, C.; Harrington, G.; and Mcinnis, M. 2014. Ecologically Valid Long-term Mood Monitoring of Individualswith Bipolar Disorder using Speech. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 4858–4862.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1746–1751.

Li, J.; Sun, Y.; Johnson, R. J.; Sciaky, D.; Wei, C.; Leaman, R.; Davis, A. P.; Mattingly, C. J.; Wiegers, T. C.; and Lu, Z. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database J. Biol. Databases Curation* .

Li, X.; Chen, Y.; Li, L.; Gao, J.; and Çelikyilmaz, A. 2017a. End-to-End Task-Completion Neural Dialogue Systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 733–743.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017b. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 986–995.

Lin, C.-Y.; and Hovy, E. 2002. From Single to Multi-document Summarization. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 457–464.

Liu, Z.; Winata, G. I.; Xu, P.; and Fung, P. 2020. Coach: A Coarse-to-Fine Approach for Cross-domain Slot Filling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 19–25.

Misra, A.; Anand, P.; Tree, J. E. F.; and Walker, M. 2015. Using Summarization to Discover Argument Facets in Online Idealogical Dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, 430–440.

Peng, N.; Poon, H.; Quirk, C.; Toutanova, K.; and Yih, W. 2017. Cross-Sentence N-ary Relation Extraction with Graph LSTMs. *Transactions of the Association for Computational Linguistics* 5: 101–115.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (Volume 1: Long Papers)*, 527–536.

Quirk, C.; and Poon, H. 2017. Distant Supervision for Relation Extraction beyond the Sentence Boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1171–1182.

Reis, H. T.; and Sprecher, S. 2009. *Encyclopedia of human relationships*. Sage Publications.

Riedel, S.; Yao, L.; and McCallum, A. 2010. Modeling Relations and Their Mentions without Labeled Text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 148–163. Springer.

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, volume 31.

Swampillai, K.; and Stevenson, M. 2010. Inter-sentential Relations in Information Extraction Corpora. In *Proceedings of the International Conference on Language Resources and Evaluation*.

Ultes, S.; Rojas-Barahona, L. M.; Su, P.; Vandyke, D.; Kim, D.; Casanueva, I.; Budzianowski, P.; Mrksic, N.; Wen, T.; Gasic, M.; and Young, S. J. 2017. PyDial: A Multi-domain Statistical Dialogue System Toolkit. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 73–78.

Verga, P.; Strubell, E.; and McCallum, A. 2018. Simultaneously Self-Attending to All Mentions for Full-Abstract Biological Relation Extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 872–884.

Yao, Y.; Ye, D.; Li, P.; Han, X.; Lin, Y.; Liu, Z.; Liu, Z.; Huang, L.; Zhou, J.; and Sun, M. 2019. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, 764–777.

Yin, Z.; Chen, F.; Ruiz, N.; and Ambikairajah, E. 2008. Speech-based cognitive load monitoring system. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2041–2044.

Yu, D.; Sun, K.; Cardie, C.; and Yu, D. 2020. Dialogue-Based Relation Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4927–4940.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Paper)*, 2204–2213.

Zhang, Y.; Qi, P.; and Manning, C. D. 2018. Graph Convolution over Pruned Dependency Trees Improves Relation Extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2205–2215.

Zhang, Y.; Zhong, V.; Chen, D.; Angeli, G.; and Manning, C. D. 2017. Position-aware Attention and Supervised Data Improve Slot Filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 35–45.

Zhang, Z.; Han, X.; Liu, Z.; Jiang, X.; Sun, M.; and Liu, Q. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, 1441–1451.

Zheng, Y.; Zhang, R.; Huang, M.; and Mao, X. 2020. A Pre-Training Based Personalized Dialogue Generation Model with Persona-Sparse Data. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, volume 34, 9693–9700.

Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; and Xu, B. 2016. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 207–212.