Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 610–614
Marseille, 11–16 May 2020

MPDD: A Multi-Party Dialogue Dataset for Analysis of Emotions and Interpersonal Relationships

Yi-Ting Chen[1], **Hen-Hsen Huang**[2,3], **Hsin-Hsi Chen**[1,3]
[1]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
[2]Department of Computer Science, National Chengchi University, Taipei, Taiwan
[3]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taipei, Taiwan
ytchen@nlg.csie.ntu.edu.tw, hhhuang@nccu.edu.tw, hhchen@ntu.edu.tw

Abstract

A dialogue dataset is an indispensable resource for building a dialogue system. Additional information like emotions and interpersonal relationships labeled on conversations enables the system to capture the emotion flow of the participants in the dialogue. However, there is no publicly available Chinese dialogue dataset with emotion and relation labels. In this paper, we collect the conversions from TV series scripts, and annotate emotion and interpersonal relationship labels on each utterance. This dataset contains 25,548 utterances from 4,142 dialogues. We also set up some experiments to observe the effects of the responded utterance on the current utterance, and the correlation between emotion and relation types in emotion and relation classification tasks.

Keywords: Multi-party dialogue, Interpersonal relationships, emotions, dataset

1. Introduction

The social dialogue system and the task-oriented dialogue system are the two major types of dialogue systems. Unlike the latter, which needs to focus on confirming user's purpose and usually targets a specific domain, the former is built in the open domain. The social dialogue system aims to understand both users' intents and their feelings, and creates an appropriate response to the users.

To design a dialogue system that understands the users' feelings, a dataset with the labels of the users' mental state is indispensable. However, those researches about the emotion detection are usually based on the datasets crawled from comment (Feng et al, 2010) and social media (Sun et al., 2010). There is few publicly available Chinese dataset containing conversation content and emotion labels similar to EmotionLines (Chen et al., 2018).

In addition to the users' emotion, we aim to investigate other factors that may affect the conversation. Interpersonal relationship between a speaker and a listener plays an important role. When people talk to friends, the wording would be much different from that talking to a stranger. If a dialogue system can predict which relation type the wording of a user belongs to, the system will be able to generate a better response.

In this paper, we release a dataset, Multi-Party Dialogue Dataset (MPDD),[1] with both emotion and relation labels on each utterance. All the dialogues are collected from TV series scripts. To the best of our knowledge, this dataset is very rare Chinese multi-party dialogue dataset with both emotion and relation labels. We also make some experiments to observe the effects of the context on the emotion and relation classification, and the correlation between emotion and relation types.

2. Related Work

Several dialogue datasets have been released in recent years. Some focus on the specific domains (Petukhova et al., 2014; Feng et al., 2016; Wei et al., 2018), and some on the multiple domains (Budzianowski et al., 2018). These datasets are annotated with the dialogue act information, and appropriate for the slot filling task when building tasked-oriented dialogue systems. In contrast, the social dialogue systems usually operate in the open domain, and need to capture users' feelings.

To train the open domain conversation model, researchers usually use the movie subtitles datasets like OpenSubtitle (Tiedemann, 2009), or use the dataset crawled from social media platforms likes Weibo (Wang et al., 2013) and Twitter (Ritter et al., 2011). However, the former has no emotion labels, which are important for training, and the latter extracts dialogues from the post-reply pairs, which are quite different from human conversation.

DailyDialog (Li et al., 2017) is a dataset in which the daily dialogues are crawled from various English learning websites and annotated with the emotion and intension type on every utterance. EmotionLines (Chen et al., 2018) is a dialogue dataset whose materials are crawled from Friends TV scripts and private Facebook messenger dialogues, and has the emotion labels on each utterance. Both DailyDialog and EmotionLines are in English. We cannot directly translate these datasets to Chinese for system development because of translation errors and the culture difference. In this paper, we aim to build a Chinese dialogue dataset with speakers' emotions and interpersonal relationships.

3. Multi-Party Dialogue Dataset (MPDD)

The source data and the annotation scheme are introduced in Section 3.1. The detail annotation process is described in Section 3.2. Section 3.3 shows the statistics of our dataset.

3.1 Dataset Development

To collect the Chinese dialogues that are close to the daily life, we crawled five TV series scripts from `www.juben108.com`, a website hosting numerous Chinese scripts. The crawled scripts are separated and each scene is regarded as a dialogue. We remove all scene descriptions and parenthetical lines, and also delete the dialogues that contain only one participant.

[1] http://nlg.csie.ntu.edu.tw/nlpresource/MPDD/

Field	Seniority	Relationship	%	Field	Seniority	Relationship	%
Family	elder	parent	7.41	Company	elder	boss	5.81
		parent-in-law	0.58		peer	colleague	7.10
		grandparent	0.36			partner	1.19
		other superior	1.11		junior	subordinate	5.47
	peer	spouse	9.66	Others	peer	couple	6.50
		brothers and sisters	5.77			friend	25.44
		other peer	2.29			enemy	3.05
	junior	child	7.31			consignor	2.10
		son/daughter-in-law	0.59			consignee	2.08
		grandchild	0.36			stranger	3.25
		other inferior	1.13			unknown	0.07
School	elder	teacher	0. 31				
	peer	classmate	0.79				
	junior	student	0. 27				

Table 1: Categories of interpersonal relationships and the percentage of each relation type.

	Neutral	Surprise	Happiness	Angry	Fear	Disgust	Sadness
%	33.19	20.16	19.34	9.91	7.43	5.02	4.95

Table 2: Distribution of emotion types in the multi-party dialogue dataset.

The dialogues includes three types of labels - emotion, relation and targeted listener labels. The emotion type is selected from Ekman's (1987) seven basic emotions, i.e., angry, sadness, fear, disgust, happiness, surprise, and neutral. That represents the speaker's emotion in the utterance. By observing the plot of all the scripts, we divide the relations between characters into 24 types shown in Table 1. These relations can also be classified from two perspectives—say, the social field and the seniority. In the classes of the social field, there are 4 classes, including family, company, school, and others. In the classes of the seniority, interpersonal relation types are separated into 3 classes, i.e., elder, peer, and junior.

3.2 Annotation Procedure

Each utterance in MPDD was annotated by three annotators. The labels with two votes are set as the ground truth. Those utterances without agreement would be re-judged by an additional annotator. The annotation of the dialogues includes four steps: noisy filtering, emotion labeling, targeted listener labeling and relation labeling. An easy and efficient annotation platform is developed and a screenshot is shown in Figure 1.

Noisy filtering: Because the dialogues were crawled from the scripts which have no standard format, there might be some noisy utterances like scene descriptions and actor's action instruction. Annotators need to mark all utterances that are not a part of the conversations to make sure the dataset is clean and sensible.

Emotion labelling: Annotators would label one emotion type from seven emotions, i.e., angry, sadness, fear, disgust, happiness, surprise, and neutral to represent the speaker's emotion in the utterance. Since the seven categories might be too vague, annotators were suggested to reference the emotion wheel[2] which provides more detailed categories for annotators to select the most appropriate type.

Targeted listener labeling: There might be more than one targeted listener in an utterance, annotators need to mark all of them. If no specific targeted listener is in the utterance, annotators regard the utterance as a declaration and mark all the participants in the dialogue as listeners.

Relation labeling: Annotators select the interpersonal relation types of all speaker-listener pairs in the dialogue, according to Table 1. If the speaker-listener pair has more than one relation at the same time, annotators need to select the one close to the dialogue situation. For example, the listener is *father* and *boss* of the speaker, and the situation of the conversation is that they are talking about the business in the company, the selected relation should be *boss*.

3.3 Dataset Statistics

Table 1 shows the percentage of each interpersonal relation type. The most common relation type is *friend*, which occupies 25.44%. Table 2 shows the percentage of each emotion type. Neutral, surprise, and happiness are the top three common types in the dataset. This distribution is similar to EmotionLines (Chen et al., 2018). It is in line with De Choudhury (2012) which showed positive emotions appearing more frequently than negative emotions.

Table 3 shows the statistics of the dataset. There are 4,142 dialogues and 25,548 utterances. On the average, there are 6.168 turns in a dialogue, and 2.338 participants are involved. A partial dialogue consisting of three utterances is shown in Table 4 as an example. For each utterance, speaker, content, emotion, listener, and speaker-listener

[2] http://westendcounselling.co.uk/emotions/wheel-of-emotion

Figure 1: A screenshot during annotation.

# dialogues	4,142
# utterances	25,548
Avg. # turns per dialogue	6.168
Avg. # participants per dialogue	2.338

Table 3: Statistics of the dataset.

Utterance 1	
Speaker	左母 "mother Zuo"
Content	那個憨女人有什麼值得送的，正鵬這個人也真是的！ "What is Zheng-Peng thinking? He has no need to send the silly woman home."
Emotion	disgust
Listener	左父 "father Zuo" : spouse
Utterance 2	
Speaker	左父 "father Zuo"
Content	哎喲，老婆子，你怎麼盡講那些不利於團結的話呢！他去送送他的同學也在情理之中嘛！ "Hey. My old woman. How can you say such uncoordinated words? It's reasonable for him to send his classmate home."
Emotion	surprise
Listener	左母 "mother Zuo": spouse
Utterance 3	
Speaker	左正鵬 "Zheng-Peng Zuo"
Content	爸、媽，我回來啦！ "Dad, Mom, I am back!"
Emotion	neutral
Listener	左父 "father Zuo" : child 左母 "mother Zuo": child

Table 4: An example dialogue consisting of three utterances.

relationship are listed. The interpersonal relation type is selected from the listener's perspective.

4. Analysis on Emotion and Relation

Multi-Party Dialogue Dataset (MPDD) is a dataset with both emotion and interpersonal relation labels. In this paper, we build classifiers for the emotion and relation classification tasks. Besides the current utterance, we also select the previous utterance and the utterance it responds (called a *responded utterance* hereafter) as one of the input features to see whether the performance is increased. For observing the correlation between emotion and interpersonal relation types, we also add either emotion or relation type as a feature to observe the effectiveness on the classification of the other type. There are four parts of the classifier in the task: responded utterance selector, contextualized embedding encoder, feature connector, and the output layer.

4.1 Responded Utterance Selector

In MPDD, every utterance has the labels that are the listeners of the utterance. We postulate that the closest utterance spoken by one of the listeners is the responded utterance that the current utterance responds to. For the utterances without responded utterances, we select the previous utterances as the alternatives.

4.2 Contextualized Embedding Encoder

To convert an utterance to the contextualized embedding, we adopt two kinds of encoders, i.e., CNN (Kim, 2014) and BERT (Devlin et al., 2018).

CNN: If there is a responded utterance, we concatenate the responded utterance and the current utterance first. Then we represent the concatenation with the word embedding, which is a 300-dimensional vector pre-trained on the Wikipedia dataset with word2vec. The word embedding matrix is fed into an 1D-Convolution layer that has 64 filters with the window size in the range of 1 to 5. After that, we fed the result into an 1D-max pooling layer, and flatten

Encoder	Baseline	w/ previous utterance	w/ responded utterance
CNN	.7169	.7207	.7207
BERT	.7259	.7494	.7632

Table 5: Accuracy for seniority classification using previous utterance and responded utterance.

Encoder	Responded utterance	Neutral	Surprise	Happiness	Angry	Fear	Disgust	Sadness
CNN	w/o	.7872	.5806	.4473	.2125	.0060	.0007	.0008
	w/	.7886	.4023	.4465	.2065	.0066	.0000	.0000
BERT	w/o	.6872	.7296	.6165	.5310	.2519	.1151	.3307
	w/	.6915	.7227	.6421	.5219	.2882	.1120	.3427

Table 6: Accuracy for each emotion class in emotion classification.

Encoder		Responded utterance	Baseline	w/ relationship	w/ seniority	w/ social field
CNN		w/o	.4852	.4875	.4860	.4867
		w/	.4522	.4560	.4521	.4525
BERT		w/o	.5851	.5852	.5837	.5830
		w/	.5927	.5893	.5834	.5881

Table 7: Accuracy for emotion classification using two encoders and interpersonal relation features.

	Encoder	Responded utterance	Baseline	w/ emotion
Relationship	CNN	w/o	.3121	.3139
		w/	.3483	.3494
	BERT	w/o	.3646	.3653
		w/	.4384	.4504
Seniority	CNN	w/o	.7169	.7167
		w/	.7247	.7240
	BERT	w/o	.7259	.7268
		w/	.7662	.7398
Social Field	CNN	w/o	.5937	.5994
		w/	.6831	.6868
	BERT	w/o	.6473	.6314
		w/	.7543	.7491

Table 8: Accuracy for relation classification using emotion features.

the output from the pooling layer to obtain the contextualized embedding.

BERT: We use the pre-trained sentence encoder, BERT, as the encoder. If there is a responded utterance, we combine these two utterances in the format of the sentence pair shown in the paper (Devlin et al., 2018). The sentence pair is fed into the BERT model, and finally the contextualized embedding is generated.

4.3 Feature Connector

To observe the correlation between emotion and interpersonal relation, we add the relation features to the emotion classifier, and vice versa. The features added to the classifier are encoded in the one-hot representation and are concatenated to generate the contextualized embedding. In addition to the emotion and relation features, we also explore the seniority and social field features to see the differences.

4.4 Output Layer

In the experiments, we use a simple dense layer with the softmax function to predict the target type of the utterance. The input dimension is based on the contextualized embedding encoder and the features we used. The output dimension is the number of the target types.

5. Results and Discussion

Section 5.1 sets up the experiments, and Section 5.2 shows and discuss the results

5.1 Experimental Setup

The maximum length of the input sentence is 50. We use Adam as the optimizer. The training batch size is 32, and the learning rate is 5×10^{-5}. The number of training epochs is tuned with validation data. We perform 5-fold cross validation and adopt the average accuracy as the metric.

5.2 Experimental Results

Table 5 shows an improvement on seniority classification when using the responded utterance instead of the previous utterance. The similar results were also obtained in social field classification. For this reason, we chose the responded utterance as the additional features.

Results are shown in Tables 6, 7 and 8. The baseline model means the classifier without additional features. From these three tables, we can observe that the models with BERT encoder defeats those with the CNN encoder in all tasks.

The responded utterance brings the significant improvement in both fine-grained relation classification (i.e., 24 interpersonal relation types) and coarse-grained

relation classification (i.e., 3 types of seniorities and 4 types of social fields), but not for the emotion classification task. The reason may be that the relation type is related to both the speaker and the listener, so the responded utterance has some information about the listener.

To check the correlation between emotion and interpersonal relation, we adopt the Pearson's Chi-squared independence test. The *p*-value between emotion and seniority, and the *p*-value between emotion and social field are lower than 0.001. Based on the statistical results, emotion and interpersonal relation are dependent.

However, both classification tasks show no improvement when emotion (interpersonal relation) features are added to interpersonal relation (emotion) classification. Thus, we need to explore other methods to capture the interactions between emotion and interpersonal relation information.

6. Conclusion and Future Work

This paper presents MPDD, which contains 4,142 dialogues and 25,548 utterances. Each utterance was annotated with an emotion label, a list of listeners and an interpersonal relationship between the speaker and listener. This dataset can be used in the classification task and the research about conversation disentanglement.

We conduct an empirical study on predicting the emotion and relation types of the utterance separately, and we also analyze the correlation between emotion and interpersonal relation. In the future work, an advanced model will be explored to both kinds of information. The generation model conditioned on both emotion and interpersonal relationship will also be investigated.

7. Acknowledgements

This research was partially supported by the Ministry of Science and Technology, Taiwan, under grants MOST-106-2923-E-002-012-MY3, MOST-108-2634-F-002-008-, MOST-108-2218-E-009-051-, and MOST-109-2634-F-002-034 and by Academia Sinica, Taiwan, under grant AS-TP-107-M05.

8. Bibliographical References

Budzianowski, P., Wen, T.-H., Tseng, B. H., Casanueva, I., Ultes, S., Ramadan, O., and Gašić, M. (2018). MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016-5026.

Chen, S.-Y., Hsu, C.-C., Kuo, C.-C., Huang, T.-H. K., Ku, L. W. (2018). EmotionLines: An Emotion Corpus of Multi-Party Conversations. In *Proceedings of the Language Resources and Evaluation Conference. arXiv preprint arXiv:* 1802.08379.

Devlin, J., Chang, W.-M., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv*:1810.04805.

Choudhury, M. D., Counts, S., and Gamon, M. (2012). Not All Moods re Created Equal! Exploring Human Emotional States in Social Media. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*.

Ekman, P., Friesen, W. V., O'sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W. A., Pitcairn, T., Ricci-Bitti, P. E., Scherer, K., Tomita, M., and Tzavaras, A. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. In *Journal of personality and social psychology*, 53(4): 712-717.

Feng, M., Xiang, B., Glass, M. R., Wang, L., and Zhou. B. (2015). Applying deep learning to answer selection: A study and an open task. 2015. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 813-820.

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. In *the 2014 Conference on Empirical Methods in Natural Language Processing*.

Liu, Q. Wei, Z., Peng, Q., Dai, X., Tou, H., Chen, T., Huang, X., and Wong, K-F. (2018). Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2: 201-207.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *The 8th International Joint Conference on Natural Language Processing*.

Petukhova, V., Gropp, M., Klakow, D., Schmidt, A., Eigner, G., Topf, M., Srb, S., Motlicek, P., Potard, B., Dines, J., Deroo, O., Egeler, R., Meinz, U., and Liersch, S. (2014). The DBOX Corpus Collection of Spoken Human-Human and Human-Machine Dialogues. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*.

Ritter, A., Cherry, C., and Dolan, W. B. (2011). Data-Driven Response Generation in Social Media. In *the 2011 Conference on Empirical Methods in Natural Language Processing*. pages 583–593.

Sun, Y.-T., Chen, C.-L., Liu, C.-C., and Liu, C,-L. (2010). Sentiment Classification of Short Chinese Sentences. In *Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing*, pages 184-198.

Tiedemann, J. (2009). News from OPUS — A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*. 5: 237-248.

Wang, H., Lu, Z., Li, H., and Chen, E. (2013). A Dataset for Research on Short-Text Conversation. In *the 2013 Conference on Empirical Methods in Natural Language Processing*.

Yang, F., Peng, Q.-K., and Xu, T. (2010). Sentiment Classification for Online Comments Based on Random Network Theory. In *Acta Automatica Sinica*, 36(6): 837-844.