CPED: A Large-Scale Chinese Personalized and Emotional Dialogue Dataset for Open-Domain Conversation

Anonymous ACL submission

Abstract

Recently, the personification and empathy capabilities of dialogue systems have received extensive attention from researchers. Although it is straightforward for humans to express themselves personally and empathically, this is highly difficult for dialogue systems since training data do not provide personalities or empathy knowledge. In this paper, we propose **CPED**, a large-scale Chinese personalized and emotional dialogue dataset, which consists of multisource knowledge related to empathy and personal characteristic. This knowledge covers 13 emotions, gender, Big Five personality traits, 19 dialogue acts and other knowledge. CPED contains more than 12K dialogues of 392 speakers from 40 TV shows. We also provide several strong baselines for open-domain conversation generation. The results show that explicitly infusing personalized knowledge and emotional information improves the personification level and empathy ability of dialogue systems, but the infusion method needs to be further studied. The dataset and baselines will be released on `https://github.com/***/CPED`.

1 Introduction

Open-domain conversation systems are of great significance in the application of human-computer interaction, companionship, depression treatment, autism intervention, etc. (Zhou et al., 2018; Zhang et al., 2020; Zheng et al., 2020b). Driving dialogue systems to learn expression capabilities from a large-scale dialogue corpus, such as OpenSubtitles (Tiedemann, 2009), Ubuntu Dialogue Corpus (Lowe et al., 2015), STC (Shang et al., 2015), LCCC (Wang et al., 2020), OpenViDial (Meng et al., 2020), etc., is considered to be feasible.

However, if we want the dialogue systems to possess a good command of personification capabilities, e.g., emotional expression, personality presentation and empathetic conversation, two critical problems need to be tackled: (i) the lack of long-term stable personalities (e.g., gender, age, and Big Five), and (ii) the lack of dynamic emotions or DAs during conversation. To the best of our knowledge, dialogue generation models considering emotion and personality as prior knowledge at the same time are currently scarce since no available dialogue dataset simultaneously provides emotional information and personalities of the speakers.

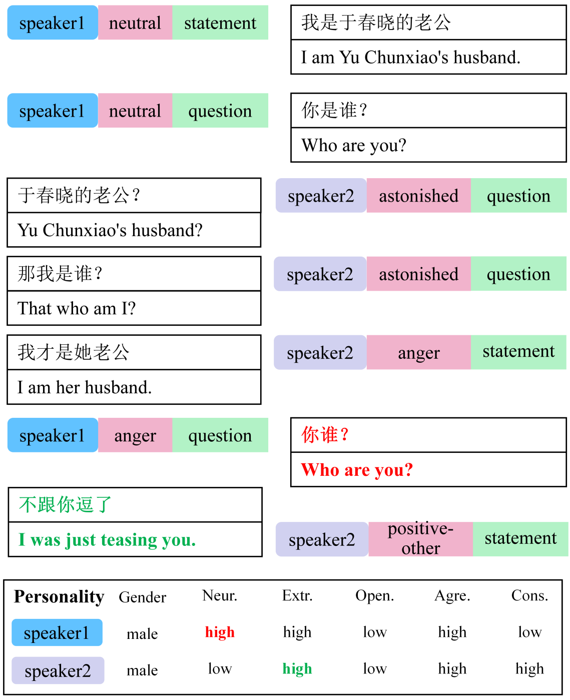

Figure 1: Example from **CPED** dataset. The dialogue consists of quadruples (speaker, emotion, DA, and utterance) along with speakers' personalities, e.g., gender, Big Five, etc. Note that the emotions or DAs of a speaker would change dynamically during conversation.

In a conversation, the participants' expression depends on not only their linguistic context but also the priori personalities and dynamic emotions. For example, in Figure 1, "speaker1" with high *neuroticism* may easily present an angry state in conversation when saying "你谁? (who are you?)". In contrast, "speaker2" with high *extraversion* and low *neuroticism*, may tend to joke during commu-

Dataset	Lang.	Modal	Dial.	Utt.	Annotation
OpenSubtitles	ML	(_,_,t)	-	11.3M	-
Twitter	EN	(_,_,t)	4,232	33K	-
Ubuntu Dialogue Corpus	EN	(_,_,t)	930K	7.1M	-
Cornell Movie Dialogs	EN	(_,_,t)	220K	304K	gender and billing-position information of characters
OpenViDial	EN	(v,_,t)	-	1.1M	-
STC	CN	(_,_,t)	4.4M	4.6M	-
Douban	CN	(_,_,t)	1.1M	6.7M	-
LCCC	CN	(_,_,t)	12M	33M	-
IEMOCAP	EN	(v,a,t)	151	7,433	10 emotions
DailyDialog	EN	(_,_,t)	13K	102K	7 emotions and 4 DAs and 10 topics
Mastodon	EN	(_,_,t)	535	2,217	3 sentiment tags and 27 DAs
MELD	EN	(v,a,t)	1,433	13,708	7 emotions
Empathetic-Dialogues	EN	(_,_,t)	25k	100K	32 emotion labels
EMOTyDA	EN	(v,a,t)	1,341	19,365	7 emotions and 12 DAs
ESTC	CN	(_,_,t)	4.4M	4.5M	6 emotions (automatically annotated)
PERSONA-CHAT	EN	(_,_,t)	10,981	164k	each personas consisting of at least 5 profile sentences
MEmoR	EN	(v,a,t)	8,536	22,732	14 emotions and 3 personality models (16PF, Big Five and MBTI)
PersonalDialog	CN	(_,_,t)	20.83M	56.25M	5 personality traits (Age, gender, location, interest, and self descriptions)
CPED(ours)	**CN**	**(v,a,t)**	**12K**	**133K**	**3 sentiments, 13 emotions, 19 DAs, 10 conversation scene, and speaker's personality (Gender, age, and Big Five)**

Table 1: Comparison among other conversation datasets and CPED. *Modal* denotes the modality of the context (*v*: video, *a*: audio, and *t*: text). *Dial.* denotes the total number of dialogues in the dataset. *Utt.* denotes the total number of utterances in the dataset. *Annotation* indicates how the dataset is labeled in terms of emotion or personality.

nication, pretending to be Yu Chunxiao's husband to joke with "speaker1". In other words, relying solely on textual contexts is insufficient to model this dialogue generation process.

Therefore, we propose a large-scale **C**hinese **P**ersonalized and **E**motional **D**ialogue dataset (**CPED**), which includes the personalities of the speakers, dynamic emotions and DAs of the multimodal dialogue contexts. CPED, which contains 12K dialogues and 133K utterances, is collected from 40 popular TV series closely related to daily life. We asked the psychology professional annotators to label the emotion and Dialogue Acts (DAs) of the speakers through video, audio and text, which is different from DailyDialog(Li et al., 2017) and ESTC(Zhou et al., 2018). In daily life, speakers may continuously speak in a round of conversation (Figure 1) during which the emotional state or DA state may change several times. Therefore, we divided a turn of dialogue into multiple utterances and annotated emotions and DAs multiple times. Furthermore, we considered gender, age and Big Five personality (BARRICK and MOUNT, 1991) as the basic personality traits.

The contributions of this paper are summarized as follows:

- We build a multiturn **C**hinese **P**ersonalized and **E**motional **D**ialogue dataset called CPED. To the best of our knowledge, CPED is the first Chinese personalized and emotional dialogue dataset. CPED contains 12K dialogues and 133K utterances with multimodal context. Therefore, it can be used in both complicated dialogue understanding and human-like conversation generation.

- CPED has been annotated with 4 types of personality knowledge (name, gender, age and Big Five personality), 2 types of dynamic emotional information (sentiment and emotion) and DAs. The personalities and emotions can be used as prior external knowledge for open-domain conversation generation, making the conversation system have a good command of personification capabilities.

- We provide baselines for personalized and emotional conversation (**PEC**), including implicit embedding and explicit infusion. This verifies the importance of using personalities and emotions as prior external knowledge for conversation generation.

2 Related Work

2.1 Open-domain Conversation Datasets

There have been various open-domain conversation datasets (Table 1(rows 2-9)) over the past few years. These datasets are usually crawled from blogs, forums, or TV series subtitle sites. OpenSubtitles (Tiedemann, 2009) is extracted from the OpenSubtitle website and includes 2.6 billion utterances across 60 languages. The Cornell Movie Dialog Corpus (Danescu-Niculescu-Mizil and Lee, 2011) involves 9,035 characters from 617 movies, including 304,713 utterances. There are also commonly used English textual conversation datasets, e.g., the Ubuntu Dialogue Corpus (Lowe et al., 2015), Twitter (Sordoni et al., 2015a) and OpenViDial (Meng et al., 2020). In the field of Chinese conversation generation, the corpus is usually crawled from social media, such as STC (Shang et al., 2015), the Douban Conversation Corpus (Wu et al., 2017) and LCCC (Wang et al., 2020). These datasets do not contain any emotional or personalized annotation information.

2.2 Emotional Conversation Datasets

Generally, the emotional perception ability of a dialogue model is defined as the task: emotion recognition in conversations (ERC) (Poria et al., 2019) or emotion reasoning (ER) (Shen et al., 2020a). Datasets, e.g., IEMOCAP (Busso et al., 2008), Mastodon (Cerisara et al., 2018), MELD (Poria et al., 2019), EMOTyDA (Saha et al., 2020), EDA (Bothe et al., 2020) and MEmoR (Shen et al., 2020a), are usually used for the ERC or ER task. These datasets generally have small sizes, with fewer than 10K dialogues, making them unsuitable for conversation generation tasks. Another type of dataset is specifically constructed for emotional conversation generation tasks (Table 1(rows 10-16)). For example, DailyDialog (Li et al., 2017) contains 13K multiturn dialogues with 102K utterances manually annotated with 7 emotions and 4 DAs. Thus, the dataset is usually used for emotional conversation generation (Zhong et al., 2019; Liang et al., 2021). EmpatheticDialogues (Rashkin et al., 2019) provides 25K dialogues with 32 types of emotion labels and 2 roles (*speaker* and *listener*) for empathetic conversation. ESTC (Zhou et al., 2018), which is annotated with six emotion categories using the Bi-LSTM emotion classifier based on the STC dataset, is used for Chinese emotional conversation generation. Unfortunately, there is no available Chinese multimodal emotional dialogue dataset so far.

2.3 Personalized Conversation Datasets

There are already some datasets related to personalized conversation (in Table 1(rows 17-19)). For example, PERSONA-CHAT (Zhang et al., 2018) crowdsourced a set of 1,155 personas and obtained 10,981 dialogs with 164,356 utterances from Turkers assigned a random persona that were asked to chat with others. PersonalDialog (Zheng et al., 2020a), a Chinese personalized conversation dataset, provides 56.25M utterances from 8.47M speakers who are annotated with personality traits, e.g., age, gender, location, interest tags, etc. MEmoR (Shen et al., 2020a), a recent multimodal emotion reasoning dataset used for the task of multimodal emotion reasoning, provides a multimodal conversation context, 14 fine-grained emotions and 3 types of personalities (16PF, Big Five and MBTI).

With explicit personality and dynamic emotional information, we believe that CPED will provide novel research opportunities and conditions for Chinese open-domain conversation, especially multimodal emotional dialogue generation and personalized dialogue generation.

3 CPED Dataset

In this section, we describe the processing stage of constructing the CPED dataset.

3.1 Video Collection and Preprocessing

Video Source In the past, Chinese conversation datasets were obtained by crawling textual dialogues from the Internet. It is difficult to obtain

# of annos.	Labels	Num.
Sentiment	*positive, neutral, and negative*	3
Emotion	*happy, grateful, relaxed, other-positive, neutral, angry, sad, feared, depressed, disgusted, astonished, worried and other-negative*	13
Gender	*male, female, and unknown*	3
Age group	*children, teenager, young, middle-aged, elderly and unknown*	6
Big Five	*high, low, and unknown*	3
DA	*greeting (g), question (q), answer (ans), statement-opinion (sv), statement-non-opinion (sd), apology (fa), command (c), agreement/acceptance (aa), disagreement (dag), acknowledge (a), appreciation (ba), interjection (ij), conventional-closing (fc), thanking (ft), quotation (ˆq), reject(rj), irony (ir), comfort (cf) and other (oth)*	19
Scene	*home, office, school, mall, hospital, restaurant, sports-venue, entertainment-venue, car, outdoor and other-scene*	11

Table 2: Annotation labels of the proposed dataset.

multimodal dialogue data and annotate the emotions and personalities based on multimodal contexts. Therefore, we searched for 100 Chinese TV series closely related to daily life and finally selected 40 TV series that had abundant emotional interaction content and sufficient characters with distinctive personalities.

Dialogue Segment Selection We built a Windows application and designed a three-step filtering process to reduce the difficulty of video selection and promote the quality of dialogue segments. Each worker was asked to learn the filtering rules and pass an assessment on which they obtained at least a 98% pass rate in the premarking stage. First, each worker was asked to watch the video and mark the start time and end time of each potential dialogue sample through the developed application. Then, whether every potential dialogue sample was suitable for CPED would be confirmed by another worker. Finally, we split the videos into dialogue segments through the video editing tool *MoviePy*[1].

Subtitle Exaction For most TV series, subtitles are embedded in videos and need to be transcribed to text using the optical character recognition (OCR) technique. We use the video OCR tool *HTWCore*[2] to generate the subtitles of each dialogue segment. Thus, we obtain the dialogue segments and their subtitles to annotate the emotions, DAs, and personalities.

[1] https://github.com/Zulko/moviepy

[2] https://github.com/xiaopinggai-webrtc/HTWCore

3.2 Annotation Scheme

Annotation Label In order for the dialogue system to learn emotional expression and personalized expression abilities, we provide multiple types of annotation labels listed in Table 2: sentiments, emotions, personalities (gender, age group and Big Five), DAs and scenes. We consider "*positive, neutral, and negative*" as the sentiment labels that are the same as MELD(Poria et al., 2019). In general, the emotion labels of conversation datasets are considered from among Ekman's six basic emotions (*joy, sad, feared, angry, surprise, and disgusted*) (Ekman et al., 1987). However, the latest studies, e.g., 32 emotion labels in EmpatheticDialogues (Rashkin et al., 2019) and 14 emotion labels in MEmoR (Shen et al., 2020b), show that more fine-grained emotion annotation can contribute to research on emotional reasoning and empathetic conversation. Considering the diversity of emotional tags and the similarity of different tags, we selected 13 emotion labels referring to EmpatheticDialogues (Rashkin et al., 2019) and 19 DA labels referring to the SWBD-DAMSL tag-set (Jurafsky et al., 1997) based on the characteristics of Chinese open-domain conversation. In particular, we have added two special labels, "*other-positive*" and "*other-negative*", which allow uncommon emotions to be included. Personality is complex and changeable, and there is no unified trait set of personality. Different from PERSONA-CHAT (Zhang et al., 2018) and PersonalDialog (Zheng et al., 2020a), we consider gender, age and Big Five personality (BARRICK and MOUNT, 1991) as the basic personality traits. Following (Li et al., 2017), we

Utterance	Speaker
多大的事你知道的我把握不好尺度	胡一菲
Big deal.You know, I can't hold the scale.	Hu Yifei
多大的事啊	胡一菲
Big deal.	Hu Yifei
你知道的我把握不好尺度	陆展博
You know, I can't hold the scale.	Lu Zhanbo

Table 3: Example of utterance overlap that need to be cut into multiple utterances correctly.

Statistics	Train	Valid	Test
# of modalities	(v,a,t)	(v,a,t)	(v,a,t)
# of TV plays	26	5	9
# of dialogues	8,086	934	2,815
# of utterances	94,187	11,137	27,438
# of speakers	273	38	81
Avg. # utt. per dial.	11.6	11.9	9.7
Max # utt. per dial.	75	31	34
Avg. emot. per dial.	2.8	3.4	3.2
Avg. DAs per dial.	3.6	3.7	3.2
Avg. utt. length	8.3	8.2	8.3
Max utt. length	127	42	45
Avg. duration	2.1s	2.12s	2.21s

Table 4: Summary of CPED dataset statistics. *utt.*, *dial.*, *emot.* refer to utterance, dialogue, emotion. (v,a,t)=(visual, audio, text).

label each dialogue as one of ten dialogue scene categories.

Annotation Process The annotation process is divided into two stages: (1) utterance-level annotation and (2) speaker-level annotation. First, we ask annotators to label the sentiments, emotions, DAs and scenes of each utterance. Second, when the dialogue samples of a TV series have been annotated, the experts are asked to annotate the *gender*, *age group* and *Big Five* of each character that appears in the dialogue samples.

3.3 Annotation Quality Control

To guarantee quality, we recruit three psychology experts who have a wealth of prior knowledge and experience for discriminating emotion, DA and personality. We jointly formulated labeling rules and labeling examples and randomly selected 200 samples for 3 rounds of prelabeling, thereby reducing the discrepancy in labeling by discussing and improving the annotation scheme. Following (Poria et al., 2019), experts are required to annotate utterances with multi-modal information that combines video, facial expressions, audio and text, which can help improve the emotional annotation accuracy. Each utterance was annotated by 3 experts, and the majority rule was used to determine the final labels. If the labeling results of the three experts are inconsistent, they needed to reannotate those utterances to find a "common" annotation. Finally, samples that still could not be labeled uniformly were discarded. In addition, since some speakers rarely speak, they will be uniformly defined as "其他 (other)", of which the gender, age group, and Big Five personality will be annotated as "unknown". Finally, we include a total of 11,835 dialogues with multi-source knowledge.

Utterance Overlap Processing Automatic subtitle extraction will be accompanied by *utterance overlap*, which means that one utterance contains the content of two speakers talking (Table 3). The statistics indicated that there were 4,613 *utterance overlaps* identified by annotators during the construction of the entire dataset. These utterance samples were correctly cut into multiple utterances, and the emotions and DAs were respectively reannotated.

3.4 Corpus Exploration

Dataset Split We randomly split the CPED dataset into three sets: train, valid and test according to the ratio of 7:1:2. In order to avoid data leakage, the split of the dataset is based on TV series, which ensures that the speakers in the training set will not appear in the valid/test set.

Dataset Statistics Figure 2 presents the distribution of the genders, ages groups, sentiments, emotions and DAs of the CPED dataset. The ratio of males to females is close to 1:1, which makes the distribution of personality and emotion close to the real world. Similar to other conversation datasets, the distribution of emotion and DA labels are unbalanced. Among them, "neutral" accounts for 32.4% of all emotions. The statistics of CPED are listed in Table 4. The average numbers of emotions per dialogue, i.e., the number of different emotion categories, are 2.8, 3.4 and 3.2 in training/validation/testing samples. The average DAs per dialogue are 3.6, 3.7, and 3.2 in training/validation/testing samples.

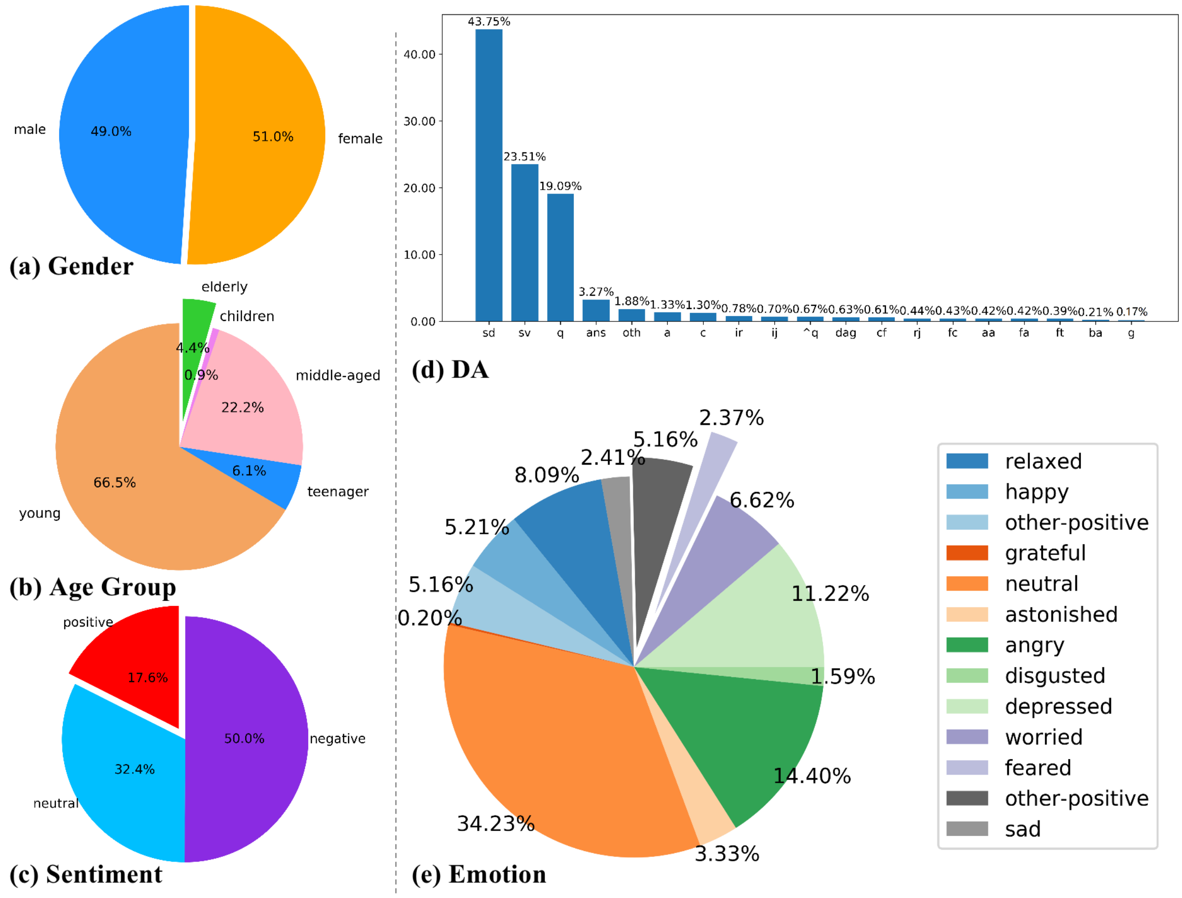

Figure 2: Distribution of Gender, Age Group, Sentiment, Emotion and DA in CPED Dataset.

4 Personalized and Emotional Conversation

In this section, we provide several benchmarks for the **P**ersonalized and **E**motional **C**onversation (**PEC**) task on the proposed CPED. Conversation generation models can usually be divided into **retrieval-based** (Yan et al., 2016; Gu et al., 2020) and **generative** (Sordoni et al., 2015b; Zhang et al., 2020; Zheng et al., 2020b). As shown in Figure 3, generative conversation models can be divided into three types: (1) w/o control signal (Luo et al., 2018; Zhang et al., 2020), (2) implicit embedding (Zheng et al., 2020b; Zandie and Mahoor, 2020; Zheng et al., 2021), and (3) explicit fusion (Zhou et al., 2018; Liang et al., 2021). Generally, the latter two architectures are used for personalized conversation generation or emotional conversation generation.

4.1 Task Definition

We research enabling the conversation generation system to generate more anthropomorphic reply content by infusing emotion and personality at the same time. **P**ersonalized and **E**motional **C**onversation (**PEC**) is defined as follows: Given the personalized information (P_{R1} and P_{R2}) of two speakers, their conversation context C, the emotion E_K and DA D_K of the response to be generated, and the personalized information P_K of the responder, the goal is to generate an anthropomorphic response Y.

$$Y = \arg\max_{Y'} P(Y'|C, E_K, D_K, P_K) \quad (1)$$

Particularly, context $C = \{(U_1, E_1, D_1, P_1), ..., (U_{K-1}, E_{K-1}, D_{K-1}, P_{K-1})\}$ contains multi-turn conversation content (i.e., utterance U_i), emotion E_i of the associated utterance, DA D_i of the associated utterance, and personalized information P_i of the associated speaker.

4.2 Baseline Models

As shown in Figure 3, we compare several categories of generative models and our method on CPED:

w/o control signal: (1) **Seq2Seq**(Sutskever et al., 2014), the classical dialogue generation model we

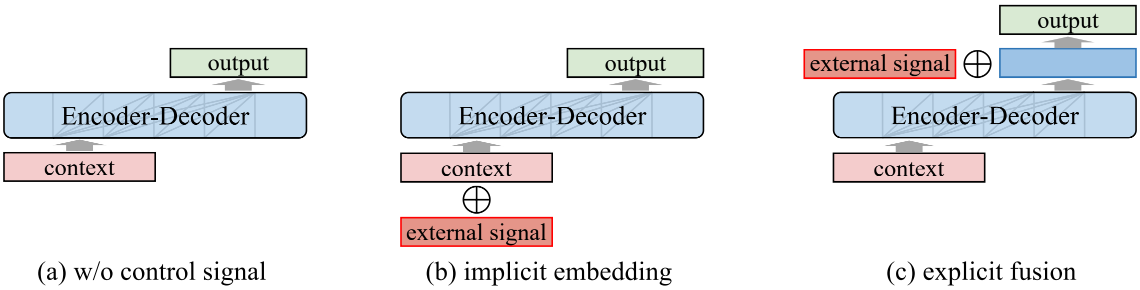

Figure 3: The generic framework of PEC. Three type of generative dialogue generation model are devised. *External signal* represents emotion, personality, DA and other prior knowledge that is used to control the conversation generation.

selected, is widely used in conversation generation. (2) **Transformer**(Vaswani et al., 2017), the second model that we evaluate, is an encoder-decoder framework based on a self-attention mechanism. The transformer has been widely applied in machine translation(Vaswani et al., 2017), language modeling(Devlin et al., 2019), dialogue generation, etc. (3) **GPT**(Zhang et al., 2020) has recently gradually been used in the field of dialog generation(Zhang et al., 2020; Wang et al., 2020). Following (Wang et al., 2020), we fine-tune CDial-GPT on the CPED dataset.

Implicit embedding: **{emo+da}-GPT** is the proposed method inspired by (Zheng et al., 2020b) that adds word embeddings E_w, segmentation embeddings E_{seq}, position embeddings E_{pos}, emotion embeddings E_{emo} and DA embeddings E_{da} together as the input embeddings for GPT:

$$E = E_w + E_{emo} + E_{da} + E_{pos} + E_{seq} \quad (2)$$

Explicit fusion: **GPT-{per+emo+da}** is the proposed method that infuses emotion E_K and DA D_K of the response to be generated and the personalized information P_K of the responder. For the emotion and DA, we constructed the embedding matrix separately to obtain emotion embedding E_g and DA embedding D_g, respectively. The embedding of personalized information is computed by a two-layer fully connected feed-forward neural network FNN(*) to project P_K to word embedding space P_g as follows:

$$P_g = FNN(P_K) \quad (3)$$

Subsequently, emotion embedding E_g, DA embedding D_g and personalization embedding P_g are concatenated together and then infused by a fully connected feed-forward neural network FNN(*) to generate control vector C_g:

$$C_g = FNN([E_g; D_g; P_g]) \quad (4)$$

We design a conditional layer to control the text generation:

$$O^c = O + g \odot C_g + (1 - g) \odot R_g \quad (5)$$

where O is the output of the last hidden layer of the language model (transformer or GPT, etc.). R_g denotes the role of the responder, which is the word embedding of “[speaker1]” or “[speaker2]”.$\odot$ is elementwise multiplication. $g \in [0, 1]$ denotes the condition weight as follows:

$$g = \sigma(FNN([O; C_g; R_g]) \quad (6)$$

where $\sigma(*)$ is an activation function (e.g., $Tanh(*)$).

4.3 Automatic Evaluation

Metrics The perplexity (**PPL**) and **BLEU** (Papineni et al., 2002) are used to evaluate the relevance and fluency of the generated responses, respectively. Then, distinct-n (**D-1**, **D-2**) (Li et al., 2016) is applied to evaluate the degree of diversity. Greedy matching (**Gre.**), embedding average (**Avg.**) (Liu et al., 2016) and F_{BERT} of BERTscore (**BERT.**) (Zhang* et al., 2020) are used to evaluate the semantic-level relevance of the generated responses and the reference responses.

Results The results in Table 5 show that it is better to explicitly infuse the emotions and personalities of the response to be generated into the conversation model than implicitly embed them. Compared to the baseline model GPT, GPT-emo achieves the best PPL (2.59↓), D-1 (0.0132↑) and

Type	Methods	Automatic.							Manual.		
		PPL	BLEU	D-1	D-2	Gre.	Avg.	BERT.	Con.	Emo.	Per.
w/o control signal	Seq2seq	107.3	0.0077	0.0252	0.1846	0.4529	0.5074	0.5196	0.823	0.726	0.684
	Transformer	62.82	**0.1680**	0.0264	0.2031	0.4674	0.5190	0.5519	1.015	0.873	0.706
	GPT	20.07	0.1171	0.0482	0.2738	0.4922	0.5509	0.5629	1.118	0.963	0.760
implicit embedding	{emo+da}-GPT	21.60	0.1304	0.0476	0.2785	0.4962	0.5552	0.5674	1.193	1.068	0.893
	w/o emo	22.84	0.1252	0.0451	0.2746	0.4964	0.5564	0.5666	1.050	0.977	0.793
	w/o da	22.09	0.1272	0.0473	0.2790	0.4962	0.5556	0.5669	1.093	0.971	0.782
explicit fusion	GPT-{emo}	**17.48**	0.1342	**0.0614**	**0.3430**	0.4996	0.5588	0.5709	1.295	1.195	0.940
	GPT-{per}	18.08	0.1372	0.0592	0.3363	0.5009	0.5606	0.5715	1.308	1.042	1.043
	GPT-{da}	17.72	0.1325	0.0605	0.3389	0.5017	0.5610	0.5703	1.285	1.047	1.003
	GPT-{per+emo}	17.70	0.1403	0.0602	0.3388	**0.5026**	**0.5617**	0.5719	1.307	**1.298**	1.075
	GPT-{per+emo+da}	17.80	0.1382	0.0601	0.3404	0.5012	0.5608	**0.5722**	**1.390**	1.232	**1.237**

Table 5: Evaluation results on CPED. The automatic evaluation includes the perplexity (**PPL**), **BLEU**, distinct-n (**D-1**, **D-2**), greedy matching (**Gre.**), embedding average (**Avg.**) and BERTscore (**BERT.**). The manual evaluation includes the content consistency (**Con.**), emotion correlation (**Emo.**) and personification capabilities (**Per.**).

D-2 (0.0692↑); GPT-{per+emo} achieves the best Gre. (0.0104↑) and Avg. (0.0108↑); and GPT-{per+emo+da} achieves the best BERT. (0.0093↑). The results demonstrate the superiority and effectiveness of explicitly infusing emotions and personalities into open-domain conversation generation.

4.4 Manual Evaluation

Metrics Three individual experts majoring in *Chinese language and literature* were asked to evaluate the generated responses in terms of content consistency (**Con.**), emotion correlation (**Emo.**) and personification capabilities (**Per.**). **Con.** denotes the consistency of the topic and content according to the conversation context. **Emo.** denotes the emotional relevance and rationality of the response generated by the dialogue system. **Per.** denotes the personification capabilities of the dialogue system and is applied to measure the human-like expression ability. The rating scale is $(0, 1, 2)$, where 0 means the worst and 2 means the best.

Results Two hundred dialogues were randomly sampled from the test set of CPED for manual evaluation. Fleiss' kappa(Fleiss, 1971) is calculated to measure the inter-rater consistency for **Con.**, **Emo.**, and **Per.**, which are 0.658, 0.632 and 0.646, indicating substantial annotation agreement respectively. Table 5 shows the results of the manual evaluation in terms of content, emotion and personification. We observe that GPT-{per+emo+da} achieves the best **Con.** (0.272↑) and the best **Per.** (0.477↑) compared with GPT while GPT-{per+emo} achieves the best **Emo.** (0.335↑). This demonstrates that "explicit fusion" can effectively benefit the conversation generation model to generate more anthropomorphic responses. Furthermore, explicitly specifying the emotion and personality of the responses will improve the emotional expression ability and personality expression ability of the dialogue system.

5 Conclusion and Future Work

In this paper, we proposed the dataset CPED, a large-scale Chinese personalized and emotional dialogue dataset containing more than 11K dialogues with 392 speakers from 40 TV shows. CPED contains abundant prior information about emotions, personalities, dialog acts and other items. The evaluation results of the baseline models are initial but indicative. Explicitly infusing emotions, personalities and dialog acts of the response to be generated can improve the personification level and emotional expression of a dialogue system. We believe that CPED can help researchers study personalized and emotional conversation (PEC).

Based on the abundant emotions, personalities, and multimodal contexts of CPED, future work can explore the following: (i) modeling or recognition of speakers' personality and emotion, (ii) prediction of responded emotion and personality, (iii) personalized and emotional conversation generation using multimodal contexts, (iv) pretrained PEC model for empathetic conversation or mental health support, etc.

References

MURRAY R. BARRICK and MICHAEL K. MOUNT. 1991. The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44(1):1–26.

Chandrakant Bothe, Cornelius Weber, Sven Magg, and Stefan Wermter. 2020. EDA: Enriching emotional dialogue acts using an ensemble of neural annotators. In *LREC*.

Carlos Busso, Murtaza Bulut, Chi Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa T. Le. 2018. Multi-task dialog act and sentiment recognition on mastodon. *CoRR*, abs/1807.05013.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Paul Ekman, Wallace V. Friesen, Maureen O"Sullivan, Anthony Chan, and Et Al. 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality & Social Psychology*, 53(4):712–717.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 2041–2044, New York, NY, USA. Association for Computing Machinery.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical Report 97-02, University of Colorado, Boulder Institute of Cognitive Science, Boulder, CO.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of The 8th International Joint Conference on Natural Language Processing (IJCNLP 2017)*.

Yunlong Liang, Fandong Meng, Ying Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. Infusing multi-source knowledge with heterogeneous graph neural network for emotional conversation generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13343–13352.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Liangchen Luo, Jingjing Xu, Junyang Lin, Qi Zeng, and Xu Sun. 2018. An auto-encoder matching model for learning utterance-level semantic dependency in dialogue generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 702–707, Brussels, Belgium. Association for Computational Linguistics.

Yuxian Meng, Shuhe Wang, Qinghong Han, Xiaofei Sun, Fei Wu, Rui Yan, and Jiwei Li. 2020. OpenViDial: A large-scale, open-domain dialogue dataset with visual contexts.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Towards emotion-aided multi-modal dialogue act classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4361–4372, Online. Association for Computational Linguistics.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586, Beijing, China. Association for Computational Linguistics.

Guangyao Shen, Xin Wang, Xuguang Duan, Hongzhi Li, and Wenwu Zhu. 2020a. Memor: A dataset for multimodal emotion reasoning in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 493–502, New York, NY, USA. Association for Computing Machinery.

Guangyao Shen, Xin Wang, Xuguang Duan, Hongzhi Li, and Wenwu Zhu. 2020b. Memor: A dataset for multimodal emotion reasoning in videos. In *Proceedings of the 28th ACM international conference on Multimedia*, pages 493–502. ACM.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015a. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado. Association for Computational Linguistics.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015b. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, volume 65, page 3104–3112.

J. Tiedemann. 2009. News from opus — a collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing V: Selected papers from RANLP 2007*, volume 5, page 237–248. Advances in Natural Language Processing.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume abs/1706.03762 of *NIPS'17*, page 6000–6010.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *CCF International Conference on Natural Language Processing and Chinese Computing(NLPCC2020)*, pages 91–103.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Vancouver, Canada. Association for Computational Linguistics.

Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 55–64, New York, NY, USA. Association for Computing Machinery.

Rohola Zandie and Mohammad H. Mahoor. 2020. Emptransfo: A multi-head transformer architecture for creating empathetic dialog systems. *CoRR*, abs/2003.02958.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 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 Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. CoMAE: A multi-factor hierarchical framework for empathetic response generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 813–824, Online. Association for Computational Linguistics.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2020a. Personalized dialogue generation with diversified traits.

Yinhe Zheng, Rongsheng Zhang, Xiaoxi Mao, and Minlie Huang. 2020b. A pre-training based personalized dialogue generation model with persona-sparse data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9693–9700.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. An affect-rich neural conversational model with biased attention and weighted cross-entropy loss. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7492–7500.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional Chatting Machine: Emotional conversation generation with internal and external memory. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pages 730–738.

A Implementation Details

We use transformers[3](Wolf et al., 2020) and CDial-GPT[4] to implement the baseline model. Emotion and DA labels are added to the dictionary as special characters through the function *add_special_tokens* of transformers for {emo+da}-GPT. The dimension of the word embeddings is set to 768, and the input length is $\leq$ 512 tokens. The dropout rate is set to 0.1, and the total number of training epochs is set to 120. We used the *AdamW* optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and the *Noam* learning rate scheduler (Vaswani et al., 2017) with $warmup\_steps = 10000$. We conduct experiments on Ubuntu 18.04 with 2 GeForce RTX 2080ti GPUs. The number of parameters in the models used and GPU hours are shown in Table 6.

[3] https://github.com/huggingface/transformers

[4] https://github.com/thu-coai/CDial-GPT

Type	Model	Param.	GPU hours
w/o control signal	GPT	95.500M	10h56m
implicitly embedding	{emo+da}-GPT	95.525M	11h25m
	w/o emo	95.515M	11h16m
	w/o da	95.510M	11h31m
explicitly fusion	GPT-{emo}	97.281M	11h21m
	GPT-{per}	97.309M	11h23m
	GPT-{da}	97.286M	11h2m
	GPT-{per+emo}	97.320M	11h27m
	GPT-{per+emo+da}	99.104M	11h36m

Table 6: Parameters and GPU hours of the models.

B Ethical Considerations

Data and Privacy All the dialogue materials are based on TV dramas (publicly available source: Tencent Video[5], Youku Video[6], iQiyi Video[7]) in which the names of the characters are all fictitious. Correspondingly, the personalities are also marked from the performance of the characters in the TV dramas.

Type	Model	Neg.	Dan.
w/o control signal	GPT	1.0%	0.5%
implicitly embedding	{emo+da}-GPT	3.5%	0.0%
	w/o emo	1.5%	0.0%
	w/o da	3.0%	0.5%
explicitly fusion	GPT-{emo}	4.5%	0.5%
	GPT-{per}	3.5%	0.5%
	GPT-{da}	0.5%	0.0%
	GPT-{per+emo}	3.5%	1.0%
	GPT-{per+emo+da}	2.5%	1.5%

Table 7: Statistics of the negative responses and dangerous responses generated by the baseline models. **Neg.** is the proportion of negative responses, and **Dan.** is the proportion of angry responses.

Potential bias and Ethical Risk We realize that if the model learns anthropomorphic expression ability, it may also learn the negative expressions or dangerous expressions brought about by personality. *Negative responses* represent those responses that make the emotions of both sides of the conversation develop in a worse direction. *Dangerous responses* represent those types of responses that

[5] https://v.qq.com

[6] https://youku.com

[7] https://iqiyi.com

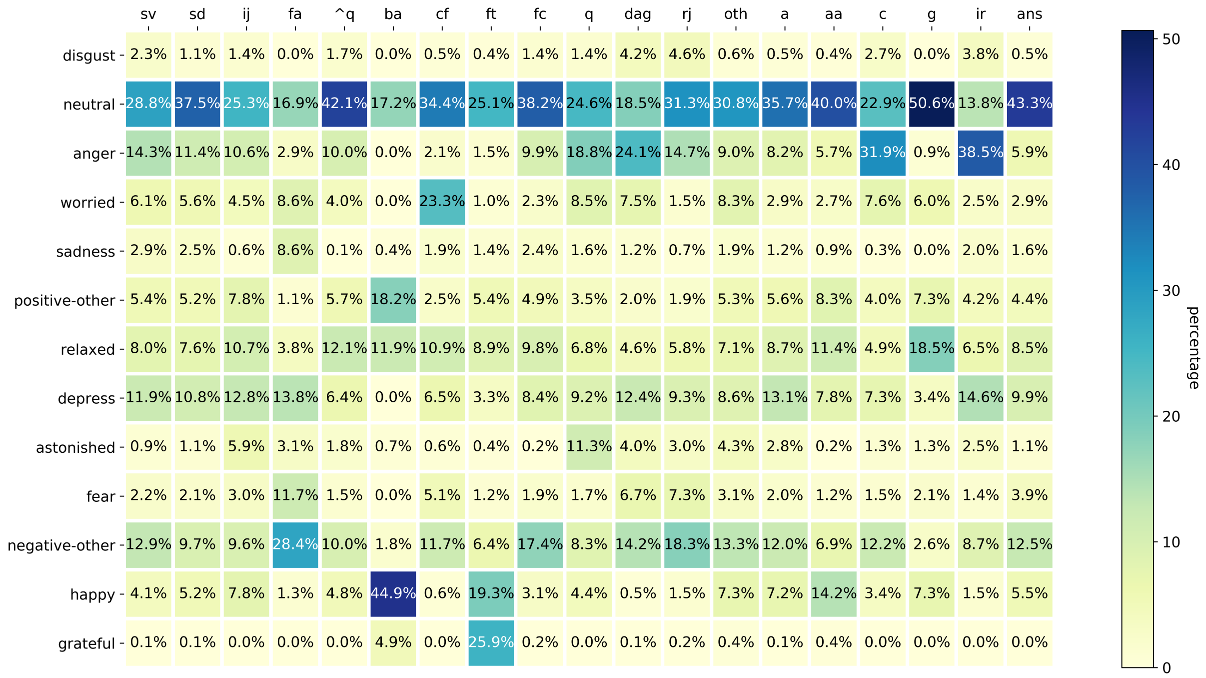

Figure 4: Relation between the Emotions and DAs.

involve suicide, abetting others to commit suicide, intimidation, etc. As shown in Table 7, we randomly selected 200 samples from the test set and counted the proportions of *negative responses* and *dangerous responses*. It is foreseeable that by improving the personification level of the dialogue generation model, it is also possible for the dialogue model to learn those risk responses. When using the CPED dataset, users should consider how to reduce the possibility of risk responses from the dialogue system while improving the level of personification of the dialogue system.

One utterance	
Dialogue_ID	01_000
Utterance_ID	01_000_000
Speaker	童文洁(Tong Wenjie)
Gender	female
Age	middle-aged
Sentiment	neutral
Emotion	neutral
Big Five	(high, high, low, low, high)
DA	greeting
Scene	other-venue
Utterance	真巧(What a coincidence)

Table 8: CPED dataset format for an utterance. Big Five = (neuroticism, extraversion, openness, agreeableness, and conscientiousness)

C Dataset sample

Each sample in the CPED dataset is composed of a series of utterance-level videos, textual context and multiple annotation results (name, gender, age group, Big Five personality, sentiment, emotion and DA). Table 8 shows the final format of one utterance on the CPED dataset in which researchers can obtain the audio file and video file corresponding to the utterance through *Utterance_ID*.

D Relationships between Emotions and DAs

Furthermore, we observed the relationships between emotions and DAs using Eq (7), as shown in Figure 4. According to the statistics, most DAs will appear at the same time as "*neutral*". "*Appreciation (ba)*" is mainly related to "*happy*" (44.9%). "*Thanking (ft)*" has an obvious correlation with "*happy*" and "*grateful*". "*Disagreement (dag)*", "*command (c)*" and "*irony (ir)*" have significant correlations with "*angry*". "*Comfort (cf)*" has an obvious correlation with "*worried*".

$$P(e|da) \approx f(e|da) = \frac{sum(e|da)}{sum(da)} \tag{7}$$

E Case Study

In Table 9, we present an example of the answers generated by the baseline models to give insight

Ex. 1	Speaker	Emotion	DA	Utterance
Dialogue History	Zeng Xiaoxian	neutral	sd	要挽留美嘉,(To retain Meijia,)
	Zeng Xiaoxian	positive(o)	sd	当然是用子乔的照片了。(Certainly, I used Ziqiao's photo.)
	Hu Yifei	anger	sd	你盗用我的创意。(You theft of my idea.)
	Zeng Xiaoxian	neutral	sd	奇怪了。(It's weird.)
	Zeng Xiaoxian	neutral	sd	我们是在比谁更了解美嘉的心思。(We are comparing who knows the mind of Meijia better.)
	Zeng Xiaoxian	neutral	sd	又不是比创意。(Not comparing the creativity.)
	Zeng Xiaoxian	relaxed	sd	要放当然放我的。(Of course, place mine.)
	Hu Yifei	negative(o)	c	把我这张也放进去。(Place mine into there too.)
	Zeng Xiaoxian	negative(o)	dag	不行。(No way.)
	Zeng Xiaoxian	depress	q	那么有效的话算谁的?(If it is effective, Whose credit is this?)
	Hu Yifei	anger	sv	当然算我的，是我的主意。(Of course it counts as mine, this is my idea.)
Golden	Zeng Xiaoxian	anger	dag	没门！(No way!)
w/o control signal	GPT	-	-	我还是想要你的。(I still want yours.)
implicitly embedding	{emo+da}-GPT	-	-	我就不用你的。(I will not use yours.)
	w/o emo	-	-	你是我的。(You belong to me.)
	w/o da	-	-	你在哪？(Where are you?)
explicitly fusion	GPT-{emo}	anger	-	我决定了。(I've made a decision.)
	GPT-{per}	-	-	怎么算？(How to judge?)
	GPT-{da}	-	dag	行吗？(Will that work?)
	GPT-{per+emo}	anger	-	我一定会好好考虑。(I will definitely think about it.)
	GPT-{per+emo+da}	anger	dag	你想得美。(When pigs fly!)

Table 9: Sample responses generated by the baseline models. The personality of the responder is (male, high, high, high, high, low) in terms of (Gender, Neuroticism, Extraversion, Openness, Agreeableness, Conscientiousness).sd: statement-non-opinion, c: command, dag: disagreement, q: question, sv: statement-opinion.

into whether the emotion and personality of the generated responses are expressed appropriately. The table shows that **GPT-{per+emo+da}** can generate highly anthropomorphic responses (e.g., 你想得美。(When pigs fly!)) with appropriate emotion and personality while the **GPT** could not express the emotion "*anger*" with the generated response "我还是想要你的。(I still want yours.)". In other words, when the emotion and DA of a response are generated and the personalities of the responder are explicitly infused into the conversation generation model, the model can perform with a high personification level and suitable emotional expression.

F Annotation Tool

We built two Windows applications for dialogue segment and annotation by using the *PyQt*[8] tool, as shown in Figure 5 and Figure 6. In the dialogue segment cutting stage, the annotators click the button "打开视频 (open video)", select an original video (about 40min), and then mark the start time and end time of the dialogue segment by repeatedly clicking the buttons "对话开始 (start of dialogue)" and " 对话结束 (end of dialogue)".

[8]https://www.riverbankcomputing.com/software/pyqt

Figure 5: Tools for dialogue segment selection.

As shown in Figure 6, annotators click "open video" to open a short dialogue video and the corresponding subtitle file. For each sentence, annotators need to select the sentiment, emotion and dialogue act. Meanwhile, they need to fill in the

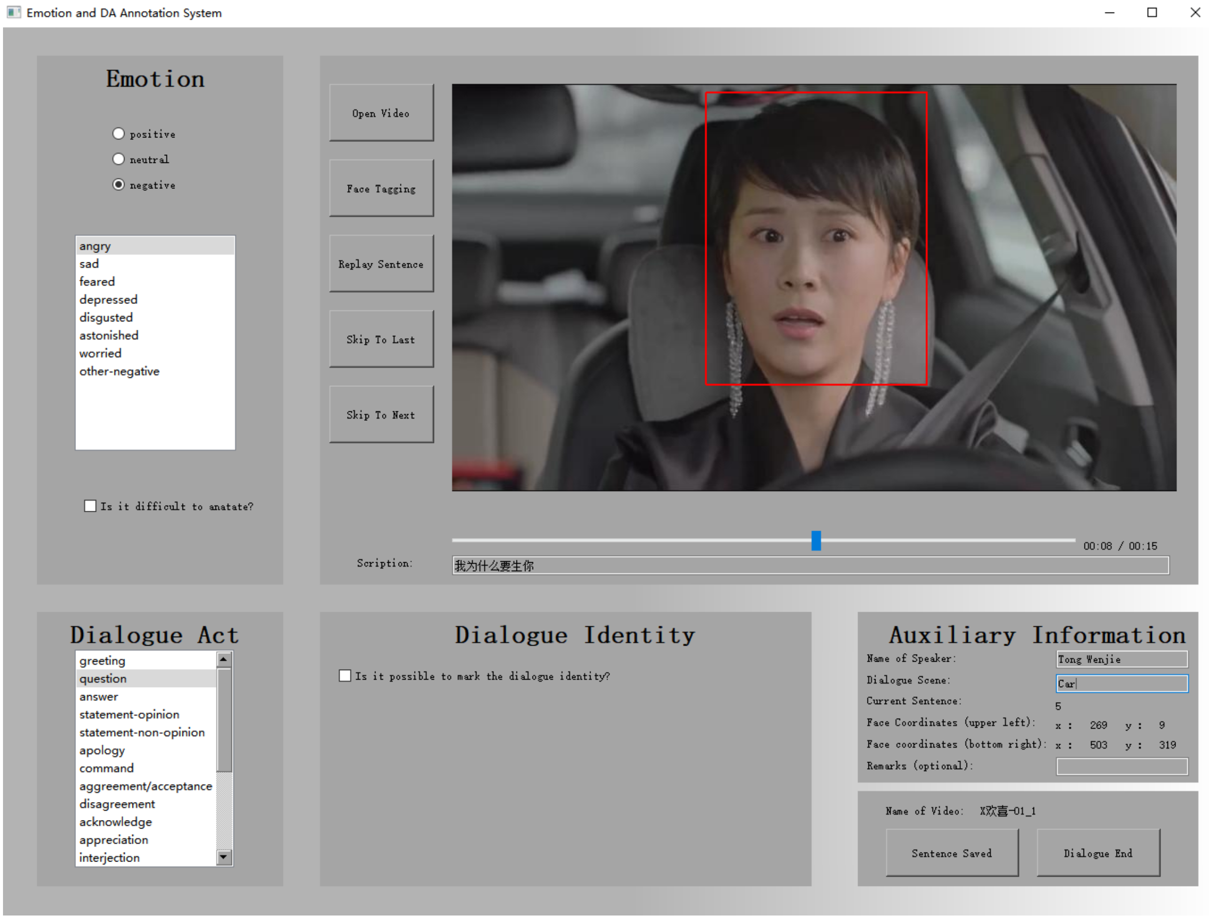

Figure 6: Conversation annotation application.

speaker's name of each sentence and the scene of the whole dialogue sample.