CPED: A Large-Scale Chinese Personalized and Emotional Dialogue Dataset for Conversational AI

Yirong Chen 1, Weiquan Fan 2, Xiaofen Xing 2, Jianxin Pang 2, Minlie Huang 2, Wenjing Han 2, Qianfeng Tie 2, and Xiangmin Xu 2

1South China University of Technology
2Affiliation not available

October 30, 2023

Abstract

Human language expression is based on the subjective construal of the situation instead of the objective truth conditions, which means that speakers’ personalities and emotions after cognitive processing have an important influence on conversation. However, most existing datasets for conversational AI ignore human personalities and emotions, or only consider part of them. It’s difficult for dialogue systems to understand speakers’ personalities and emotions although large-scale pre-training language models have been widely used. In order to consider both personalities and emotions in the process of conversation generation, we propose CPED, a large-scale Chinese personalized and emotional dialogue dataset, which consists of multi-source knowledge related to empathy and personal characteristic. These knowledge covers gender, Big Five personality traits, 13 emotions, 19 dialogue acts and 10 scenes. CPED contains more than 12K dialogues of 392 speakers from 40 TV shows. We release the textual dataset with audio features and video features according to the copyright claims, privacy issues, terms of service of video platforms. We provide detailed description of the CPED construction process and introduce three tasks for conversational AI, including personality recognition, emotion recognition in conversations as well as personalized and emotional conversation generation. Finally, we provide baseline systems for these tasks and consider the function of speakers’ personalities and emotions on conversation.

Our motivation is to propose a dataset to be widely adopted by the NLP community as a new open benchmark for conversational AI research. The full dataset is available at https://github.com/scutcyr/CPED.
CPED: A Large-Scale Chinese Personalized and Emotional Dialogue Dataset for Conversational AI

Yirong Chen, Student Member, IEEE, Weiquan Fan, Student Member, IEEE, Xiaofen Xing, Member, IEEE, Jianxin Pang, Minlie Huang, Senior Member, IEEE, Wenjing Han, Qianfeng Tie, Xiangmin Xu, Senior Member, IEEE

Abstract—Human language expression is based on the subjective construal of the situation instead of the objective truth conditions, which means that speakers’ personalities and emotions after cognitive processing have an important influence on conversation. However, most existing datasets for conversational AI ignore human personalities and emotions, or only consider part of them. It’s difficult for dialogue systems to understand speakers’ personalities and emotions although large-scale pretraining language models have been widely used. In order to consider both personalities and emotions in the process of conversation generation, we propose CPED, a large-scale Chinese personalized and emotional dialogue dataset, which consists of multi-source knowledge related to empathy and personal characteristic. These knowledge covers gender, Big Five personality traits, 13 emotions, 19 dialogue acts and 10 scenes. CPED contains more than 12K dialogues of 392 speakers from 40 TV shows. We release the textual dataset with audio features and video features according to the copyright claims, privacy issues, terms of service of video platforms. We provide detailed description of the CPED construction process and introduce three tasks for conversational AI, including personality recognition, emotion recognition in conversations as well as personalized and emotional conversation generation. Finally, we provide baseline systems for these tasks and consider the function of speakers’ personalities and emotions on conversation. Our motivation is to propose a dataset to be widely adopted by the NLP community as a new open benchmark for conversational AI research. The full dataset is available.

Index Terms—Dialogue system, cognitive processing, conversation generation, data collection.

I. INTRODUCTION

OPEN-DOMAIN conversation systems are of great significance in the application of human-computer interaction, companionship, depression treatment, autism intervention, etc. [1]–[3]. Driving dialogue systems to learn expression capabilities from a large-scale dialogue corpus, such as OpenSubtitles [4], Ubuntu Dialogue Corpus [5], STC [6], LCCC [7], OpenViDial [8], etc., is considered to be feasible.

However, if we expect 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 dialogue acts (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.

In a conversation, the participants’ expression depends not only on 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 con-
trast, "speaker2" with high extraversion and low neuroticism, may tend to joke during communication, pretending to be Yu Chunxiao’s husband to joke with "speaker1". People’s personality is imperceptibly affecting their own expression style. In other words, relying solely on supervised learning on textual contexts is insufficient to model this dialogue generation process. Besides, according to the book Cognitive psychology: Applying the science of the mind [9], there are also significant differences in Conversation Styles between female and male speakers. On the whole, only providing large-scale text for training conversation generation models can not make them master human cognitive expression patterns.

Therefore, we propose a large-scale Chinese Personalized and Emotional Dialogue 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, making its distribution of personality or emotion close to the real world. We asked the psychology professional annotators to label the emotions and DAs of the speakers through video, audio and text, which is different from DailyDialog [10] and ESTC [1]. 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 [11] as the basic personality traits.

The contributions of this paper are summarized as follows:

1) We build a multiturn Chinese Personalized and Emotional Dialogue 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 multi-modal context. Therefore, it can be used in both complicated dialogue understanding and human-like conversation generation.

2) CPED has been annotated with 3 character attributes (name, gender age), Big Five personality traits, 2 types of dynamic emotional information (sentiment and emotion) and DAs. The personality traits 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.

3) We propose three tasks for CPED: personality recognition in conversations (PRC), emotion recognition in conversations (ERC), and personalized and emotional conversation (PEC). A set of experiments verify the importance of using personalities and emotions as prior external knowledge for conversation generation.

The remainder of this paper is organized as follows: Section II discusses the related work; we then describe the construction process and detailed characteristics of CPED in Section III; definition of personality and emotion recognition in conversations, and corresponding baseline experiments are elaborated in Section IV; definition and baseline experiments of personalized and emotional conversation are described in Section V; applications and limitations of CPED are presented in Section VI; finally, Section VII illustrates the conclusion and future work.

II. RELATED WORK

A. Cognitive Psychology Theory for Conversation

a) Personality Theory: Allport proposed the personality traits [12] in 1921. Allport claims that personality trait has the ability to dominate individual behavior, and divides personality traits into two categories: common traits and individual traits. Cattell proposed the Sixteen Personality Factor Questionnaire (16PF) in 1949 [13]. Eysenck proposed the structure of personality [14] in 1953, and developed the Eysenck Personality Questionnaire (EPQ) [15] in 1975. Early lexical studies on personality models [16], [17] have proved that the terms used to describe personality traits in English are mainly composed of five dimensions, that is named the five factor personality model. McCrae & Costa (1997) established a five factor personality model based on 16PF factor analysis, which are Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness [18]. They also released a NEO Personality Inventory (NEO-PI) [19] in 1992 and NEO-PI-R [18] in 1997. Typical five factor personality inventories include Hogan Personality Inventory (HPI) [20] and Big Five Inventory (BFI) [21]. Tellegen & Waller (1987) [22] used the method of random stratified sampling to select 400 adjectives for self description, and then did factor analysis to obtain the seven dimensions of personality, and put forward the Big Seven factor model of of personality, which are Positive Emotionality (PEM), Negative Valence (NVAL), Positive Valence (PVAL), Negative Emotionality (NEM), Dependability (DEP), Agreeableness (AGR), Conventionality (CONV). Among multifarious personality models, the Big Five(BF, also called OCEAN) [18], [21] personality model has been proved to have cross-cultural applicability and has been widely used. In addition, Marusic & Bratko (1998) found that different gender has different distribution in each personality dimension of BF [23]. Soto et al. (2011) found that BF personality domains have mean-level age differences [24]. Therefore, gender, age group and BF of the speakers are taken into account in the annotation label.

b) Emotion Theory: More than 90 definitions of “emotion” have been proposed in the past according to Plutchik’s research [25]. Izard (1991) divides emotions into three parts: subjective experience, external performance and physiological arousal [26]. At present, there are two basic representative formats for the classification of emotions: dimensional emotional state (DES) and categorical emotion states (CES) [27]. Russell (1980) defines emotion as two continuous scales: valence and arousal [28]. The valence–arousal–dominance space (VAD) [29] and the pleasure-arousal-dominance space (PAD) [30] are the commonly used DES models, which transform the complex emotions into continuous 3D space. However, it is very difficult to annotate the continuous emotional labels, which consumes a lot of time and human resources, especially for the text. The CES models hold that emotions have completely different structures. Tomkins (1970) believes that there are eight primary emotions [31].
Izard (1991) put forward that there are 10 basic emotions [26]. In 1980, Robert proposed the Plutchik’s emotions (happiness, surprise, anger, disgust, fear, sadness, anticipation, joy, surprise, and trust) [33]. Izard (1991) put forward that there are 10 basic emotions [26].

### B. Conversation Datasets

In Table I, we briefly review the available conversation datasets.

#### a) Open-domain Conversation Datasets: There have been various open-domain conversation datasets (Table I(rows 2-10)) over the past few years. These datasets are usually crawled from blogs, forums, or TV series subtitle sites, e.g., OpenSubtitles [4], Cornell Movie Dialog Corpus [35], Ubuntu Dialogue Corpus [5], Twitter [34] and OpenViDial [8]. In the field of Chinese conversation generation, the corpus is usually crawled from social media, such as STC [6], the Douban Conversation Corpus [36], LCCC [7] and WDC-Dialogue [37]. Among them, WDC-Dialogue [37] has 1.4 billion dialogues so that the pre-training model can be fully trained in the field of open-domain dialogue generation. These datasets do not contain any emotional or personalized annotation information. Therefore, the dialogue generation model (e.g. DialoGPT [2], CDialGPT [7]) can only learn personalized or emotional expressions through the dialogue context (single-modal or multi-modal) provided by the corpus.

#### b) Emotional Conversation Datasets: Generally, the emotional perception ability of a dialogue model is defined as the task: emotion recognition in conversations (ERC) [40] or emotion reasoning (ER) [46]. Datasets, e.g., IEMOCAP [38], Mastodon [39], MELD [40], EMOTyDA [42], EDA [50], MEmoR [46] and M²ED [43], 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. For example, DailyDialog [10] contains 13K multi-turn dialogues with 102K utterances manually annotated with 7 emotions and 4 DAs. Thus, the dataset is usually used for emotional conversation generation. Unfortunately, there is no available large-scale Chinese multimodal emotional dialogue dataset for emotional conversation generation so far.

#### c) Personalized Conversation Datasets: There are already some datasets related to personalized conversation (Table I(rows 19-24)). For example, PERSONA-CHAT [44] crowdsourced a set of 1,155 personas and obtained 10,981 dialogues with 164,356 utterances from Turkers assigned a random persona that were asked to chat with others. In particular, each persona consists of at least 5 profile sentences, just like a

| Dataset | Lang. | Modal | Dial. | Utt. | Annotation |
|---------|-------|-------|-------|------|------------|
| OpenSubtitles [4] | ML | (v,a,t) | - | 11.3M | - |
| Twitter [34] | EN | (v,a,t) | 4.232 | 33K | - |
| Ubuntu Dialogue Corpus [5] | EN | (v,a,t) | 930K | 7.1M | - |
| Cornell Movie Dialogs [35] | EN | (v,a,t) | 220K | 304K | gender and billing-position information of characters |
| OpenViDial [8] | EN | (v,a,t) | - | 1.1M | - |
| STC [6] | CN | (v,a,t) | 4.4M | 4.6M | - |
| Douban [36] | CN | (v,a,t) | 1.1M | 6.7M | - |
| LCCC [7] | CN | (v,a,t) | 12M | 33M | - |
| WDC-Dialogue [37] | CN | (v,a,t) | 1.4B | 3.0B | - |
| IEMOCAP [38] | EN | (v,a,t) | 151 | 7,433 | 10 emotions |
| DailyDialog [10] | EN | (v,a,t) | 13K | 102K | 7 emotions and 10 topics |
| Mastodon [39] | EN | (v,a,t) | 535 | 2,217 | 3 sentiment tags and 27 DAs |
| MELD [40] | EN | (v,a,t) | 1,433 | 13,708 | 7 emotions |
| EmpatheticDialogues [41] | EN | (v,a,t) | 25k | 100K | 32 emotion labels |
| EMOTyDA [42] | EN | (v,a,t) | 1.341 | 19,365 | 7 emotions and 12 DAs |
| ESTC [1] | CN | (v,a,t) | 4.4M | 4.5M | 6 emotions (automatically annotated) |
| M²ED [43] | CN | (v,a,t) | 990 | 24,449 | 7 emotions, role names, ages and genders |
| PERSONA-CHAT [44] | EN | (v,a,t) | 10,981 | 164k | each personas consisting of at least 5 profile sentences |
| PEC [45] | EN | (v,a,t) | 355K | 833K | persona sentences for empathetic conversations from subreddits happy and offmychest |
| MEmoR [46] | EN (v,a,t) | 8,536 | 22,732 | 14 emotions and 3 personality models (16PF, Big Five and MBTI) |
| FriendsPersona [47] | EN (v,a,t) | 711 | 8,157 | Big Five personality traits of speaker in a dialogue |
| PELD [48] | EN (v,a,t) | 6,510 | 10,468 | 7 emotions and Big Five personality traits |
| PersonalDialog [49] | CN (v,a,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) |
A. Video Collection and Preprocessing

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

b) 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.

c) Subtitle Extraction: 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 to generate the subtitles of each dialogue segment. Thus, we obtain the dialogue segments and their subtitles to annotate the emotions, DAs, and personalities.

III. CPED DATASET

In this section, we describe the processing stage of constructing the CPED dataset. To construct a Chinese personalized and emotional dialogue dataset, we collected a large number of TV series related to daily life, and asked the crowdworkers to filter the dialogue segments with abundant emotions and personalities. These dialogue segments were annotated in terms of emotions and personalities by 3 full-time staff of psychology major. In the following, we describe each processing stage of constructing CPED dataset: (1) collecting and preprocessing videos, (2) designing the annotation labels, (3) annotating the emotions and personalities, (4) ensuring annotation quality and re-annotating the overlapping utterance segments.

A. Video Collection and Preprocessing

a) Video Source: In the past, 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. Specifically, PERSONA-CHAT and PersonalDialog provide actually character attributes rather than personality traits. FriendsPersona is annotated with BF personality traits of speakers, which is used for personality recognition on multi-party dialogues. However, the BF personality traits of speaker in FriendsPersona change in different conversations, which is contradictory to the personality coherence. PELD is proposed for predicting emotion for response using BF personality traits and VAD vector, in which the personality traits are averaged with personality traits of FriendsPersona. MEmoR, 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, BF and MBTI). MEmoR is mainly used for the task of multimodal emotion reasoning, in which the personalities are used for improving the performance of emotion reasoning. At present, in the field of Chinese conversation, there is a lack of personality related datasets, which hinders the research on personality related tasks, such as personality recognition in conversations.

With explicit personality and dynamic emotional information, we believe that CPED will provide novel research opportunities and conditions for Chinese open-domain conversation, e.g., personality recognition and emotion recognition on conversations, personalized and emotional conversation.

B. Annotation Scheme

a) 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 II: sentiments, emotions, personalities (gender, age group and BF), DAs, and scenes. We consider “positive, neutral, and negative” as the sentiment labels that are the same as MELD. In general, the emotion labels of conversation datasets are considered from among Ekman’s six basic emotions (joy, sad, feared, angry, surprise, and disgusted) [54]. However, the latest studies, e.g., 32 emotion labels in EmpatheticDialogues [41] and 14 emotion labels in MEmoR [55], 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 [41] and 19 DA labels referring to the SWBD-DAMSL tag-set [56] 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 [44] and PersonalDialog [49], we consider gender, age and BF personality as the basic personality traits. According to the Developmental Psychology, the age groups are divided into: children (< 11), teenager (12 – 20), young (21 – 39), middle-aged (40 – 60) and elderly (> 60). Following Dailydialog [10], we label each dialogue as one of ten dialogue scene categories.

 b) 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. In particular, the Chinese Big Five Inventory–2 (Chinese BFI-2) [57] proposed by Zhang et al. is used for calculating the scores of Big Five personalities. Annotators were asked to fill Chinese BFI-2 for each speaker. The normalized average of the final score is used to judge the personality traits (high, low, and unknown).

C. Annotation Tool

We built two Windows applications for dialogue segment and annotation by using the PyQt tool, as shown in Figure 2 and Figure 3. In the dialogue segment cutting stage, the annotators click the button "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)".

As shown in Figure 3, 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 speaker’s name of each sentence and the scene of the whole dialogue sample.

D. 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 [40], 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 “other (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.

### TABLE II

| # of annos. | Labels                                                                 | Num. |
|-------------|------------------------------------------------------------------------|------|
| Sentiment   | positive, neutral, and negative, happy, grateful, relaxed, other-positive, neutral, angry, sad, feared, depressed, disgusted, astonished, worried and other-negative | 3    |
| Emotion     | g, q, a, ans, sv, a, a, i, c, f, t, r, ir, cf and other                  | 13   |
| Gender      | male, female, and unknown                                              | 3    |
| Age group   | children, teenager, young, middle-aged, elderly                       | 6    |
| Big Five    | high, low, and unknown                                                 | 3    |
| Scene       | home, office, school, mall, hospital, restaurant, sports-venue, entertainment-venue, car, outdoor and other-scene | 11   |

### TABLE III

**Example of Utterance Overlap That Need to Be Cut Into Multiple Utterances Correctly.**

| 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. | Hui Yifei |
| 多大的事啊         | 胡一菲 |
| Big deal.          | Hui Yifei |
| 你知道的我把握不好尺度 | 陆震博 |
| You know, I can’t hold the scale. | Lu Zhanbo |
a) 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 III). 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.

E. Corpus Exploration

a) 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.

b) Dataset Statistics: Figure 4 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 IV. 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. As shown in Figure 5, the proportion of high is higher than that of low in Extraversion,
Openness, Agreeableness and Conscientiousness, while lower in Neuroticism.

![Fig. 5. Distribution statistics of Big Five](image)

**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 V 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*.

**TABLE V**

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

**IV. PERSONALITY AND EMOTION RECOGNITION IN CONVERSATIONS**

We are committed to making the dialogue system acquire cognitive ability like human, including understanding the personalities of the speaker and speaker’s current emotion through conversations. Therefore, we research two subtasks respectively: personality recognition in conversations (PRC) and emotion recognition in conversations (ERC).

**A. Personality Recognition in Conversations (PRC)**

**a) Task Definition:** Given a speaker’s conversation with others, it is required to recognize the speaker’s personality traits through the conversation record, which includes two scenarios, (1) 1 − 1 conversations: the robot recognizes the personality traits of the speaker through the conversation between them (e.g., psychological counseling), (2) 1 − N conversations (see Figure 6): the robot listens to the speaker’s conversations with other N people and then recognizes the speaker’s personality traits (e.g., group chatbot, home service robot). Since 1 − N includes the case of 1 − 1, we only discusses PRC in 1 − N conversations. The task of PRC in 1 − N conversations can be formulated as:

\[
Per_i = \arg \max_{Per'_i} P(Per'_i|C_{i,j}, \cdots, C_{i,N}),
\]

where \(Per_i = [\text{Neu}, \text{Ext}, \text{Ope}, \text{Agr}, \text{Con}]\) is a 5-dimensional vector representing Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness, \(C_{i,j}\) is the conversations between \(Speaker_i\) and \(Speaker_j\) (1 ≤ j ≤ N).

**b) Baseline Models:** Several benchmark models are provided for PRC task, including

1) **BERT** only uses the concatenation of the utterances of the target \(Speaker_i\) [47] as the input to BERT [58], while **BERT** uses the full conversation \(C_{i,j}\) in the natural order. The personality \(Per_i\) of \(Speaker_i\) is calculated as follows:

\[
h_{i,j}^{[CLS]} = BERT(T_{i,j}),
\]

\[
h^{[CLS]} = [h_{i,1}^{[CLS]}, \cdots, h_{i,N}^{[CLS]}],
\]

\[
Per_{i,k} = MLP_k(\text{Avgpooling}(h^{[CLS]})),
\]

where \(T_{i,j} = U_{i,j}\) for **BERT** and \(C_{i,j}\) for **BERT**. \(U_{i,j} = [CLS], u_{1}, u_{2}, \cdots, u_{m}\) is the concatenation of the utterances of the target \(Speaker_i\) in conversation \(C_{i,j}\). \([CLS]\) is the special token of BERT. \(MLP_k\) is a multi-layer perceptron. **Avgpooling** means average pooling. \(k\) is the index of personality in different dimensions.

2) **BERT** uses the full conversation text as the input to BERT and a shared SeNet [59] as the feature fusion layer of local personality features of different conversations, which can be formulated as:

\[
Per_{i,k} = MLP_k(SeNet(h^{[CLS]})),
\]

where \(k \in [1, 5]\) is the dimension index of personalities.

3) **BERT** uses the full conversation text as the input to BERT and five independent Senets as the feature fusion layer of local personality features of different conversations, which can be formulated as:

\[
Per_{i,k} = MLP_k(SeNet_k(h^{[CLS]})),
\]
B. Emotion Recognition in Conversations (ERC)

a) Task Definition: ERC task focuses on identifying the sentiment-level or emotion-level labels $e_M$ of the utterance $u_M$ according to the conversation context $U = u_1, \cdots, u_M$:

$$e_M = \arg\max_{e'} P(e'|u_1, \cdots, u_{M-1}, u_M),$$

(7)

where $u_i$ is the utterance contains several tokens spoken by $Speaker(u_i)$. $M$ is the number of the utterances in the conversation.

b) Baseline Models: In order to provide an effective benchmark, we consider two types of benchmark models: ERC models with current single utterance as input and ERC models with current utterance and dialogue history as input.

1) ERC models with current single utterance $u_M$ as input, include TextCNN [60], TextRNN [61], TextRCNN [62], FastText [63], BERT [58]. They use sentence-level language model $LM$ to obtain the representation of $u_M$, and then use $MLP$ to predict the emotion $e_M$ as follows:

$$h = LM(u_M),$$
$$e_M = MLP(h),$$

(8)

(9)

2) ERC models with both current utterance and dialogue history include bcLSTM [64], DialogueRNN [65], DialogueGCN [66], DialogXL [67] and EmoBERTa [68].

bcLSTM is a bi-directional contextual LSTM model that has two unidirectional and opposite-direction LSTMs stacked together. DialogueRNN is a classic and efficient algorithm for ERC, which uses three gated recurrent units (GRU) to model the dialogue process, including the global GRU, the party GRU and the emotion GRU. DialogueGCN captures richer contextual information by considering the speaker information of the utterance and the relative positions of the target utterance and the context, which has three stages, consisting of sequential context encoding, speaker-level context encoding and classification. DialogXL used pretrained language models for ERC, in which the memory-saving utterance recurrence mechanism and dialog-aware self-attention are used. EmoBERTa is a speaker-aware model based on RoBERTa, which preods speaker names to utterances. We also proposed a baseline model BERT+AVG+MLP based on BERT [58], which adds all the speaker names to the special token dictionary and splices the speaker names and utterances sequentially as the input of BERT. The average pooling of hidden-layer output of BERT then inputs to multi-layer perceptron (MLP) to predict the emotion labels.

c) Experimental Results: The results of ERC in CPED is shown in Table VII. Considering only the current utterance for emotion recognition, FastText achieves the state-of-the-art performance for negative emotion while poor performance on the emotion classes neutral (24.76) and positive (0.95). Other utterance-level models have the same defects, mainly because the ability of these models to deal with label imbalance (see Figure 4(c)) is weak. The dialogue-level models can better handle the adverse effects caused by label imbalance. For example, BERT+AVG+MLP achieves the state-of-the-art performances in average accuracy and Macro-F1, since it has relatively good performance in three emotional polarities. The emotion consistence and emotion mutation also affect the performance of dialogue-level models. The probability of emotion transition is shown in Figure 7. The probability of emotion mutation in negative, neutral and positive are 0.225, 0.337 and 0.427 respectively, which makes the ERC task significantly different from other long text emotion recognition tasks. In the future, it is necessary to further study the influence and challenge of emotion consistence and emotion mutation in

| Model | Neur. | Ext. | Ope. | Agr. | Con. | Avg. |
|-------|-------|------|------|------|------|------|
| BERT$^a$ | 50.75 | 78.08 | 57.93 | 85.76 | 63.60 | 67.23 | 72.93 |
| BERT$^c$ | 55.29 | 78.08 | 53.90 | 80.98 | 63.35 | 66.32 | 72.69 |
| BERT$_{genet}^c$ | 53.40 | 77.71 | 55.42 | 81.99 | 61.59 | 66.02 | 71.89 |
| BERT$_{asenet}^c$ | 53.27 | 78.21 | 55.42 | 85.89 | 63.48 | 67.25 | 74.08 |

| Model | Neg. | Neu. | Pos. | Avg. |
|-------|------|------|------|------|
| TextCNN [60] | 64.51 | 24.56 | 14.04 | 48.90 |
| TextRNN [61] | 62.69 | 33.21 | 15.32 | 47.89 |
| TextRCNN [62] | 64.04 | 31.03 | 18.78 | 49.13 |
| FastText [63] | 65.30 | 24.76 | 0.95 | 48.62 |
| BERT [58] | 59.97 | 42.98 | 32.60 | 48.96 |
| bcLSTM [64] | 59.06 | 38.86 | 28.28 | 49.65 |
| DialogueRNN [65] | 59.06 | 36.16 | 44.97 | 48.57 |
| DialogueGCN [66] | 60.65 | 36.99 | 40.19 | 47.69 |
| EmoBERTa [68] | 59.34 | 39.02 | 41.53 | 48.09 |
| DialogXL [67] | 60.45 | 41.45 | 41.74 | 51.24 |
| BERT+AVG+MLP | 61.40 | 40.10 | 42.95 | 51.50 |

TABLE VI
PERSONALITY RECOGNITION IN CPED (NEU.: NEUROTICISM, EXT.: EXTRAVERSION, OPE.: OPENNESS, AGR.: AGREEABLENESS, CON.: CONSCIENTIOUSNESS AND AVG.: AVERAGE).

TABLE VII
SENTIMENT RECOGNITION IN CPED (NEG.: NEGATIVE, NEU.: NEUTRAL, POS.: POSITIVE AND AVG.: AVERAGE).
V. PERSONALIZED AND EMOTIONAL CONVERSATION

In this section, we provide several benchmarks for the Personalized and Emotional Conversation (PEC) task on the proposed CPED. Conversation generation models can usually be divided into retrieval-based [69], [70] and generative [2], [3], [71]. As shown in Figure 8, generative conversation models can be divided into three types: (1) w/o control signal [2], [72], (2) implicit embedding [3], [73], [74], and (3) explicit fusion [1], [52]. Generally, the latter two architectures are used for personalized conversation generation or emotional conversation generation.

A. Task Definition

We research enabling the conversation generation system to generate more anthropomorphic reply content by infusing emotion and personality at the same time. Personalized and Emotional Conversation (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.

\[ Y = \arg \max_{Y'} P(Y'|C, E_K, D_K, P_K) \]  

(10)

Particularly, context \( C = \{(U_1, E_1, D_1, P_1), \ldots, (U_{K-1}, E_{K-1}, D_{K-1}, P_{K-1})\} \) 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.

B. Baseline Models

As shown in Figure 8, we compare several categories of generative models and our method in CPED:

a) w/o Control Signal: (1) Seq2Seq [75], the classical dialogue generation model we selected, is widely used in conversation generation. (2) Transformer [76], 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 [76], language modeling [58], dialogue generation, etc. (3) GPT [2] has recently gradually been used in the field of dialog generation [2], [7]. Following [7], we fine-tune CDial-GPT on the CPED dataset.

b) Implicit Embedding: \{emo+da\}-GPT is the proposed method inspired by [3] 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:

\[ E = E_w + E_{emo} + E_{da} + E_{pos} + E_{seq} \]  

(11)

c) Explicit Fusion: \( \text{GPT-}\{\text{per+emo+da}\} \) 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 MLP(\( * \)) to project \( P_K \) to word embedding space \( P_g \) as follows:

\[ P_g = \text{MLP}(P_K) \]  

(12)

Subsequently, emotion embedding \( E_g \), DA embedding \( D_g \) and personalized embedding \( P_g \) are concatenated together and then infused by a MLP(\( * \)) to generate control vector \( C_g \):

\[ C_g = \text{MLP}([E_g; D_g; P_g]) \]  

(13)

We design a conditional layer to control the text generation:

\[ O^c = O + g \odot C_g + (1-g) \odot R_g \]  

(14)

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 element-wise multiplication. \( g \in [0,1] \) denotes the condition weight as follows:

\[ g = \sigma(\text{MLP}([O; C_g; R_g])) \]  

(15)

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

C. Implementation Details

We use transformers\(^5\) [77] and CDial-GPT \(^6\) 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 [76] with warmup_steps = 10000. We conduct experiments on Ubuntu 18.04 with 2 GeForce RTX 2080ti GPUs.

\(^5\)https://github.com/huggingface/transformers

\(^6\)https://github.com/thu-coai/CDial-GPT
D. Automatic Evaluation

a) Metrics: The perplexity (PPL) and BLEU [78] are used to evaluate the relevance and fluency of the generated responses, respectively. Then, distinct-n (D-1, D-2) [79] is applied to evaluate the degree of diversity. Greedy matching (Gre.), embedding average (Avg.) [80] and F\textsubscript{BERT} of BERTScore (BERT) [81] are used to evaluate the semantic-level relevance of the generated responses and the reference responses.

b) Results: The results in Table VIII show that it is better to explicitly fuse 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-emote achieves the best PPL (2.59, D-1 (0.0132) and 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.

E. Manual Evaluation

a) 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.

b) Results: Two hundred dialogues were randomly sampled from the test set of CPED for manual evaluation. Fleiss’ kappa [82] 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 VIII 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.

![Fig. 8. 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.](image-url)

### Table VIII

| Type                  | Method | 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 | 1.237 | 1.390 | 1.237 |
| Transformer           |        | 62.82      | \textbf{0.1680} | 0.0264  | 0.2031  | 0.4674 | 0.5190 | 0.5519 | 1.298 | 1.298 | 1.298 |
|                       | GPT    | 20.07      | 0.1171  | 0.0482  | 0.2738  | 0.4922 | 0.5509 | 0.5629 | 1.232 | 1.390 | 1.237 |
| implicit embedding    |        |            |         |         |         |       |       |       |      |      |      |
| {emo+da}-GPT w/o emo  |        | 21.60      | 0.1304  | 0.0476  | 0.2785  | 0.4962 | 0.5532 | 0.5674 | 1.193 | 1.068 | 0.893 |
| w/o da                |        | 22.84      | 0.1252  | 0.0451  | 0.2746  | 0.4964 | 0.5564 | 0.5666 | 1.268 | 1.091 | 0.812 |
| explicit fusion       |        |            |         |         |         |       |       |       |      |      |      |
| GPT-\{emo\}          |        | \textbf{17.48} | 0.1342 | \textbf{0.0614} | \textbf{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  | \textbf{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 | \textbf{0.5722} | \textbf{1.390} | 1.232 | 1.237 |
TABLE IX
SAMPLE RESPONSES GENERATED BY THE BASELINE MODELS. SD: STATEMENT-NON-OPTION, C: COMMAND, DAG: DISAGREEMENT, Q: QUESTION, SV: STATEMENT-OPTION.

| 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         | relaxed       | sd   | 又不是比创意。(Not comparing the creativity.)                            |
| Hu Yifei              | negative(o)   | c    | 要放当然放我的。(Of course, place mine.)                                 |
| Zeng Xiaoxian         | negative(o)   | dag  | 不行。(No way.)                                                           |
| Zeng Xiaoxian         | depress       | q    | 为了有效的话算谁的? (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

| w/o control signal   | GPT           |      |                                                                            |
|----------------------|---------------|------|---------------------------------------------------------------------------|
| implicitly embedding | {emo+da}-GPT  | -    | 我还是想要你的。 (I still want yours.)                                   |
| w/o emo              | -             | -    | 我就不用你的。 (I will not use yours.)                                   |
| w/o da               | -             | -    | 你在哪里? (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!)                                              |

F. Case Study

In Table IX, we present an example of the answers generated by the baseline models to give insight 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., 你想得了美嘉。(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.

VI. APPLICATIONS AND LIMITATION OF CPED

A. Applications

CPED allows evaluation of both conversational cognitive tasks and conversation generation tasks, e.g. speaker modeling, personality recognition in conversations, emotion recognition in conversations, DA recognition in conversations, emotion prediction for response, emotional conversation generation, personalized conversation generation, empathetic conversation etc. By being multimodal, CPED can also be applied in multimodal personality or emotion recognition, multimodal conversation generation. It will play a positive role in promoting the development of cognitive intelligence.

B. Ethical Considerations

a) Data and Privacy: All the dialogue materials are based on TV dramas (publicly available source: Tencent Video7, Youku Video8, iQiyi Video9) 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. The video and audio clips are licensed under Copyright Law of the People’s Republic of China. According to the privacy issues, copyright claims, terms of service of Tencent Video, Youku Video and iQiyi Video, we only release the textual dataset with audio features and video features.

b) Difference between television conversation and natural conversation: Similar to FriendsPersona [47] and MELD [40], CPED is derived from TV shows. According to the book Television Dialogue: The sitcom Friends vs natural conversation [83], television conversations and natural conversations are basically the same in terms of linguistic features. However, television conversations tend to present a limited set of scenarios, interaction types and topic categories. To this end, when we select TV series, we try to cover different scenes of daily life as much as possible (see Table II). In addition, due to the entertainment characteristics of TV shows, screenwriters often use expletive, slang, appraisal and other language means to achieve humorous effects, in order to make the language in TV shows more authentic and attractive. Therefore, the emotion distribution of television conversations

7https://tv.qq.com
8https://youku.com
9https://iqiyi.com
may be different from natural conversations (see Figure 4-(c)). When using this dataset, you can consider joint pre-training with unlabeled natural conversation datasets to alleviate this problem.

c) 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 involve suicide, abetting others to commit suicide, intimidation, etc. As shown in Table X, 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.

VII. CONCLUSION AND FUTURE WORK

In this paper, we proposed a challenging dataset CPED for conversational AI, 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. We introduce three challenging tasks for conversational AI research in CPED, including personality recognition, emotion recognition in conversations as well as personalized and emotional conversation generation. The evaluation results of the baseline models are initial but indicative. In Chinese conversations, the tasks of personality recognition and emotion recognition need to be further with linguistic characteristics and psychological knowledge, e.g. differences in different personality dimensions, differences of the demonstrated personalities or emotions in conversations between speakers and different characters, linguistic characteristics of different sentiment polarities and etc. For personalized and emotional conversation, 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 both the cognitive processing in conversations and the personalized and emotional conversation (PEC) task. 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.

TABLE X

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

REFERENCES

[1] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, “Emotional Chatting Machine: Emotional conversation generation with internal and external memory,” in Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 730–738. [Online]. Available: https://www.aaai.org/ojs/index.php/AAAI/AAAI18/paper/viewFile/16455/15753

[2] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, “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. Online: Association for Computational Linguistics, Jul. 2020, pp. 270–278. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-demos.30

[3] Y. Zheng, R. Zhang, X. Mao, and M. Huang, “A pre-training based personalized dialogue generation model with persona-sparse data,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 9693–9700. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6518

[4] J. Tiedemann, “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, vol. 5. Advances in Natural Language Processing, oct 2009, p. 237–248. [Online]. Available: https://benjamins.com/catalog/cilt.309.19tie

[5] R. Lowe, N. Pow, I. Serban, and J. Pineau, “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. Prague, Czech Republic: Association for Computational Linguistics, Sep. 2015, pp. 285–294. [Online]. Available: https://www.aclweb.org/anthology/W15-4640

[6] L. Shang, Z. Lu, and H. Li, “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). Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 1577–1586. [Online]. Available: https://www.aclweb.org/anthology/P15-1152

[7] Y. Wang, P. Ke, Y. Zheng, K. Huang, Y. Jiang, X. Zhu, and M. Huang, “A large-scale chinese short-text conversation dataset,” in CCF International Conference on Natural Language Processing and Chinese Computing(NLPCC2020), 2020, pp. 91–103. [Online]. Available: https://arxiv.org/abs/2008.03946

[8] Y. Meng, S. Wang, Q. Han, X. Sun, F. Wu, R. Yan, and J. Li, “OpenViDial: A large-scale, open-domain dialogue dataset with visual contexts,” 2020. [Online]. Available: https://arxiv.org/abs/2012.15015

[9] B. Robinson-Riegler and G. Robinson-Riegler, Cognitive psychology: Applying the science of the mind. Pearson, 2016. [Online]. Available: https://idun.augsburg.edu/monographs/35

[10] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, “DailyDialog: A manually labelled multi-turn dialogue dataset,” in Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 986–995. [Online]. Available: https://aclanthology.org/I17-1099

[11] M. R. BARRICK and M. K. MOUNT, “The big five personality dimensions and job performance: A meta-analysis,” Personnel Psychology, vol. 44, no. 1, pp. 1–26, 1991. [Online]. Available: https://doi.org/10.1111/j.1744-6570.1991.tb00688.x
A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, R. Plutchik and H. Kellerman, “Emotion: Theory, research, and H. Schlosberg, “Three dimensions of emotion.” Psychological review Journal of personality 
J. A. Russell, “A circumplex model of affect.” The psychology of emotions 
A. Tellegen and N. G. Waller, “Re-examining basic dimensions of O. P. John, “The "big five" factor taxonomy: Dimensions of personality 
R. Plutchik, “The nature of emotions: Human emotions have deep R. Mehrabian, “Pleasure-arousal-dominance: A general framework tools for clinical practice,” American scientific, vol. 89, no. 4, pp. 344–350, 2001. Online: https://www.jstor.org/stable/27887503 
C. E. Izard, The psychology of emotions. Springer Science & Business Media, 1991. 
S. Zhao, G. Jia, J. Yang, G. Ding, and K. Keutzer, “Emotion recognition from multiple modalities: Fundamentals and methodologies,” IEEE Signal Processing Magazine, vol. 38, no. 6, pp. 59–73, 2021. Online: https://doi.org/10.1109/MSP.2021.3106905 
J. S. Oso, O. P. John, S. D. Gosling, and J. Potter, “Age differences in emotionality traits from 10 to 65: Big five domains and facets in a large cross-sectional sample.” Journal of personality and social psychology, vol. 100, no. 2, pp. 330–48, 2011. Online: https://doi.org/10.1037/a0021717 
R. Plutchik, “The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice,” American scientist, vol. 89, no. 4, pp. 344–350, 2001. Online: https://www.jstor.org/stable/27887503 
C. E. Izard, The psychology of emotions. Springer Science & Business Media, 1991. 
S. Zhao, G. Jia, J. Yang, G. Ding, and K. Keutzer, “Emotion recognition from multiple modalities: Fundamentals and methodologies,” IEEE Signal Processing Magazine, vol. 38, no. 6, pp. 59–73, 2021. Online: https://doi.org/10.1109/MSP.2021.3106905 
J. A. Russell, “A circumplex model of affect.” Journal of personality and social psychology, vol. 39, no. 6, p. 1161, 1980. Online: https://doi.org/10.1037/0022-3514.39.6.1161 
H. Schlosberg, “Three dimensions of emotion.” Psychological review, vol. 61, no. 2, p. 81, 1954. Online: https://doi.org/10.1037/h0054570 
A. Mehrabian, “Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament,” Current Psychology, vol. 14, no. 4, pp. 261–292, 1996. Online: https://doi.org/10.1007/BF02686918 
S. Tomkins, “Affects as primary motivational systems,” Feelings and emotions, pp. 101–110, 1970. 
P. Ekman and W. V. Friesen, “Constants across cultures in the face and emotion.” Journal of personality and social psychology, vol. 17, no. 2, pp. 124–129, 1971. Online: https://doi.org/10.1037/0022-3514.17.2.124 
R. Plutchik and H. Kellerman, “Emotion: Theory, research, and experience,” in Theories of Emotion, R. Plutchik and H. Kellerman, Eds. Academic Press, 1980, p. ii. Online: Available: https://doi.org/10.1016/B978-0-12-558701-3.50001-6 
A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J. A. Russell, and B. De Choudhury, “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, Denver, Colorado: Association for Computational Linguistics, May–Jun. 2015, pp. 196–205. [Online]. Available: https://www.aclweb.org/anthology/N15-1020 
C. Danesucu-Niculescu-Mizil and L. Lee, “Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialog,” in Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 76–87. [Online]. Available: https://www.aclweb.org/anthology/W11-0609 
Y. Wu, W. Wu, C. Xing, M. Zhou, and Z. Li, “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). Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 496–505. [Online]. Available: https://www.aclweb.org/anthology/P17-1046 
H. Zhou, P. Ke, Z. Zhang, Y. Gu, Y. Zheng, C. Zheng, Y. Wang, C. H. Wu, H. Sun, X. Yang, B. Wen, X. Zhu, M. Huang, and J. Tang, “EVA: an open-domain Chinese dialogue system with large-scale generative pre-training,” CoRR, vol. abs/2108.01547, 2021. [Online]. Available: https://arxiv.org/abs/2108.01547 
C. Bussu, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “IEMOCAP: interactive emotional dyadic motion capture database,” Language Resources and Evaluation, vol. 42, no. 4, pp. 335–559, 2008. Online: Available: https://link.springer.com/article/10.1007/s10579-008-9076-6 
C. Cerisara, S. Jafari-Tanjani, A. Oluoskan, and H. T. Le, “Multi-task dialog act and sentiment recognition on mastodon,” CoRR, vol. abs/1807.05013, 2018. Online: http://arxiv.org/abs/1807.05013 
P. Soria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, “MELD: A multimodal multi-party dataset for emotion recognition in conversations,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 527–536. Online: https://www.aclweb.org/anthology/P19-1050 
H. Rashkin, E. M. Smith, M. Li, and Y.-L. Bourreau, “Towards empathetic open-domain models: A new benchmark and dataset,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 5370–5381. [Online]. Available: https://www.aclweb.org/anthology/P19-1534 
T. Saha, A. Patra, S. Saha, and P. Bhattacharyya, “Towards emotion-aided multi-modal dialogue act classification,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, Jul. 2020, pp. 4361–4372. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.402 
J. Zhao, T. Zhang, J. Hu, Y. Liu, Q. Jin, X. Wang, and H. Li, “M3ED: Multi-modal multi-scene multi-label emotional dialogue database,” in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 5699–5710. [Online]. Available: https://anthology.acl.org/2022.acl-long.391 
S. Zhang, E. Dinan, J. Urbaneck, A. Szlam, D. Kiela, and J. Weston, “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). Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2204–2213. [Online]. Available: https://www.aclweb.org/anthology/P18-1205 
P. Zhong, C. Zhang, H. Wang, Y. Liu, and C. Miao, “Towards persona-based empathetic conversational models,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, Nov. 2020, pp. 6556–6566. [Online]. Available: https://anthology.acl.org/2020.emnlp-main.531 
G. Shen, X. Wang, X. Duan, H. Li, and W. Zhu, “Memor: A dataset for multimodal emotion reasoning in videos,” in Proceedings of the 28th ACM International Conference on Multimedia, ser. MM ’20. New York, NY, USA: Association for Computing Machinery, 2020, p. 493–502. [Online]. Available: https://doi.org/10.1145/3394171.3413909 
J. Zhao, X. Zhang, X. Kang, J. Jiang, X. Zhang, and D. Choi, “Automatic text-based personality recognition on monologues and multiparty dialogues using attentive contextual networks and contextual embeddings (student abstract),” in Proceedings of the AAAI Conference on Artificial Intelligence, 

Furthermore, the relationships between emotions and DAs are shown in Figure 9. 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”.

APPENDIX A

RELATIONSHIPS BETWEEN EMOTIONS AND DAs

Fig. 9. Relation between the Emotions and DAs.