A Cognitive Stimulation Dialogue System with Multi-source Knowledge Fusion for Elders with Cognitive Impairment

Jiyue Jiang, Sheng Wang, Qintong Li, Lingpeng Kong, Chuan Wu
The University of Hong Kong
{jiangjy,u3009618,qtli}@connect.hku.hk
{lpk,cwu}@cs.hku.hk

Abstract

When communicating with elders with cognitive impairment, cognitive stimulation (CS) help to maintain the cognitive health of elders. Data sparsity is the main challenge in building CS-based dialogue systems, particularly in the Chinese language. To fill this gap, we construct a Chinese CS conversation (CSConv) dataset, which contains about 2.6K groups of dialogues with therapy principles and emotional support strategy labels. Making chit chat while providing emotional support is overlooked by the majority of existing cognitive dialogue systems. In this paper, we propose a multi-source knowledge fusion method for CS dialogue (CSD), to generate open-ended responses guided by the therapy principle and emotional support strategy. We first use a progressive mask method based on external knowledge to learn encoders as effective classifiers, which is the prerequisite to predict the therapy principle and emotional support strategy of the target response. Then a decoder interacts with the perceived therapy principle and emotional support strategy to generate responses. Extensive experiments conducted on the CSConv dataset demonstrate the effectiveness of the proposed method, while there is still a large space for improvement compared to human performance[1].

1 Introduction

Dialogue systems have enjoyed rapid progress in recent years, through communication with humans to satisfy diverse needs (Liu et al., 2021; Kann et al., 2022). Cognition stimulation of elders is a critical psychological therapy where dialogue systems serve as effective tools for restoring the cognition of older adults (De Oliveira et al., 2014; Park et al., 2019; Tokunaga et al., 2021).

Some studies have shown that chit-chat can help older people with cognitive restoration (van Rijn et al., 2010; Garcia, 2022). Meanwhile, several studies have shown that emotional support is beneficial for maintaining or even increasing cognitive function in elders (Ellwardt et al., 2013; Liu et al., 2020; Sharma et al., 2020). Nonetheless, there remains an open question on how to introduce emotional support and therapy principles simultaneously into chit-chat dialogue systems to provide cognitive recovery training for elders with cognitive impairment.

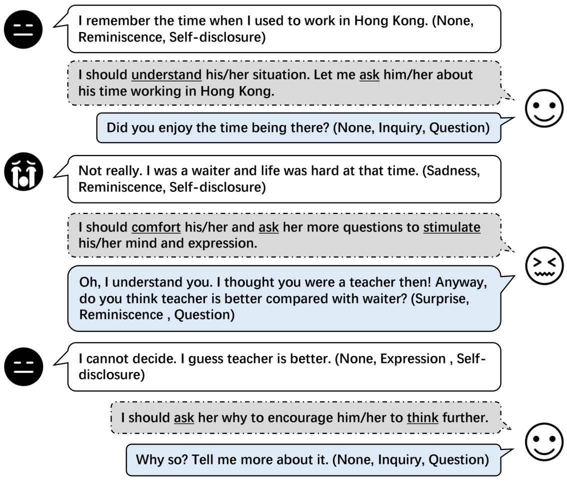
I remember the time when I used to work in Hong Kong. (None, Reminiscence, Self-disclosure)
I should understand his/her situation. Let me ask him/her about his time working in Hong Kong.
Did you enjoy the time being there? (None, Inquiry, Question)
Not really. I was a waiter and life was hard at that time. (Sadness, Reminiscence, Self-disclosure)
I should comfort his/her and ask her more questions to stimulate his/her mind and expression.
Oh, I understand you. I thought you were a teacher then! Anyway, do you think teacher is better compared with waiter? (Surprise, Reminiscence , Question)
I cannot decide. I guess teacher is better. (None, Expression , Self-disclosure)
I should ask her why to encourage him/her to think further.
Why so? Tell me more about it. (None, Inquiry, Question)

Figure 1: An example of a Chinese CS-based dialogue from the CSConv dataset (translated from Chinese to English), being provided to the elders (left) by the therapist (right). The emotion classification, therapy principle, support strategy are marked in the parentheses after the utterances. The underline highlight the emotion words and keywords.

One main obstacle to building cognitive dialogue is the lack of training corpora, especially in the Chinese language. Therefore, we first construct a Chinese **CS Conv**ersation (**CSConv**) dataset, containing about 2.6K groups of dialogue data where each utterance is annotated with three types of labels, i.e., therapy principle, emotional labels, and emotional support strategy labels. To generate open-ended responses with emotional support strategies,

[1] Our data and code could be found in https://github.com/jiangjyjy/CSD

we propose a multi-source knowledge fusion in a Chinese **CS** **D**ialogue (**CSD**) system. We use Jiagu[2], a Chinese NLP toolkit, to extract emotional words and keywords to form knowledge source and progressively mask the extracted knowledge on the encoder side, to increase the generalizability of the model. Meanwhile, we adopt Chinese EmoBank (Lee et al., 2022) to calculate the weight value of each word in the utterance, so that the model pays more attention to words with high values. By introducing multiple sources of external knowledge, we greatly enrich the content of the conversation. Moreover, we judge the content and emotions that elders express which is critical to generate satisfied responses, matching them with the cognitive therapeutic principles, and coming up with corresponding supporting strategies. At last, we design a multi-source interactive mechanism so that emotional support strategies and cognitive stimulation therapies can be reasonably combined to generate responses benefiting to mental health. Figure 1 shows an example of a conversation with an elder based on the therapy principle.

In summary, our contributions are as follows: (1) We construct a Chinese CS-based conversation dataset to facilitate the following research; (2) We propose a progressive mask method for encoder modules, which enhances the generalizability on emotional knowledge and the applicability of the therapeutic conversations with elders; (3) We design a multi-source interactive method to model the interaction among encoder modules, decoder modules and external knowledge; (4) We conduct extensive experiments to demonstrate the effectiveness of the proposed CSD.

2 Dataset

2.1 Data Collection

There is no publicly available CS-based Chinese conversation dataset to enable a cognitively restorative dialogue system for elders with cognitive impairment. We introduce a Chinese one-to-one open-domain **CS Conv**ersation dataset, (**CSConv**), which is collected and created via cognitive stimulation therapy videos and handbook[3], and the ratio of conversation data from videos to those from the handbook is approximately 2:1.

As high-quality conversation examples are needed for building Chinese CS-based dialogue system, our efforts include the following. (1) The videos are Cantonese. We first translate the Cantonese conversations in the videos into Mandarin Chinese, in a format suitable for CS model training. (2) We make Mandarin conversations artificially based on the eighteen therapy principles in the handbook. (3) We clean the dataset based on rules (e.g., truncating excessively long utterances, removing the multiple consecutive symbols in the utterance). (4) We manually annotate whether each utterance is spoken by the SPEAKER or the LISTENER (SPEAKER for elder, LISTENER for smart speaker or health care worker). (5) We use BERT-based text classification models to annotate the emotion label, strategy label, therapy label of each utterance, and then conduct manual review and modification. (6) All the data are professionally produced and reviewed twice. (7) We test our CSConv dataset on some text classification models and text generation models, which can directly reflect the performance differences between models.

Therapy Labels	**Explanation**
None	Neutral
Inquiry	Ask questions for information or open-domain questions
Respect	Be respectful or use a set pattern when talking to older people
Reminiscence	Remember things elders did when elders were a child, as well as things elders did before and personal information
Expression	Improve elders language skills and expression
Enjoyment	To have fun in conversation or to enjoy something
Comfort	Comfort the elderly to some extent

Table 1: Therapy Labels and their interpretation.

The CSConv dataset consists of about three thousand conversations, separated by blank rows. Each line in each conversation represents the utterance of SPEAKER or LISTENER, and SPEAKER and LISTENER's utterances alternate. The format of each line is: SPEAKER/LISTENER utterance + <CS> + therapy label + <EMO> + emotion label + <strategy> + strategy label, where <CS> is the separator of therapy label and SPEAKER/LISTENER utterance; <EMO> is the separator of therapy label and emotion label; <Strategy> is the separator of emotion label and strategy label. There are eight types of emotional labels, namely none, disgust, sadness, fear, surprise, like, happiness and anger. There are nine strategies (i.e., None, Question, Reflection of Feelings, Self-disclosure, Providing Sug-

[2] https://github.com/ownthink/Jiagu

[3] https://www.brainlive.socialwork.hku.hk/

gestions, Information, Others), which are similar to the strategies in (Liu et al., 2021). There are seven types of therapy labels. Table 1 shows the name of explanation of each therapy label.

2.2 Data Statistics

Statistics of the CSConv dataset are given in Table 2. The number and proportion of therapy labels, emotion labels and strategy labels are shown in Table 3.

Categories	**Number**
Conversations	2643
Utterances	16845
SPEAKER Utterances	8363
LISTENER Utterances	8480
Average token per conversation	60.39
Average utterance per conversation	6.37
Average token per utterance	9.48

Table 2: Statistics of the CSConv dataset.

Therapy Labels	**Number**	**Proportion**
None	5296	31.44
Inquiry	4156	24.67
Respect	2134	12.70
Reminiscence	464	2.76
Expression	2651	15.74
Enjoyment	1862	11.05
Comfort	281	1.67
Emotion Labels	**Number**	**Proportion**
None	12060	71.60
Disgust	273	1.62
Sadness	629	3.74
Fear	62	0.37
Surprise	355	2.11
Like	1317	7.82
Happiness	1954	11.60
Anger	193	1.15
Strategy Labels	**Number**	**Proportion**
None	7060	41.92
Question	4195	24.91
Reflection of feelings	293	17.40
Self-disclosure	3022	17.94
Providing suggestions	262	1.56
Information	819	4.86
Others	1190	7.07

Table 3: Number and proportion of therapy, emotion, strategy labels.

3 Method

3.1 Overview

Figure 2 gives an overview of our Chinese CSD architecture, which consists of two stages: (1) Progressive mask encoder; (2) Multi-source interactive decoder. The first stage is divided into two modules: progressive mask encoder for context training and encoders for text classification.

3.2 Progressive Mask Encoder

Progressive Mask Encoder for Context Training. Like the traditional BERT pre-training task, in order to better represent information of the utterances and evaluate the Next Sentence Prediction (NSP) task, the utterances of the SPEAKER and LISTENER are used to generate three types of embeddings (Vaswani et al., 2017), namely word embedding, position embedding and segment embedding.

During training, the encoder randomly masks tokens to improve generalizability. We first use Jiagu's sentiment analysis function to extract entities (i.e., one and multiple words) and sentences with positive or negative values generated by Jiagu greater than the λ_{emo} threshold, and Jiagu's keyword extraction function to extract keywords in the utterances. Eventually, emotion and keyword dictionaries are constructed. Through the emotion and keyword dictionaries, the data during training is masked in pre-defined proportions. As the training progresses, the span of a single mask gradually increases (i.e., from one word to multiple words, and finally to a sentence), the ratios of masking one-word entities, two-word entities, three-word entities, four-word entities and sentences are λ_1, λ_2, λ_3, λ_4 and λ_5, respectively. In order to further improve the encoder's generalization through the progressive mask method, we retain a certain proportion of the traditional BERT mask method. To be more specific, 5% of the entities in the utterances are randomly masked, of which 80% proceed mask processing, 10% proceed random replacement processing, and 10% remain unchanged.

After the progressive mask operation, encoders are used to encode context information for the utterances (i.e., context learning) and finally the pre-trained models are obtained.

Encoders of context training based on the emotion dictionary are used for utterance emotion classification. Encoders based on the keyword dictionary are used to classify the therapy principle and support strategy of the utterances.

Encoders for Text Classification. A multi-turn dialogue context consists of M utterances emitted by SPEAKER and LISTENER in turn. The context $\mathcal{U}$ refers to the sequence of utterance,

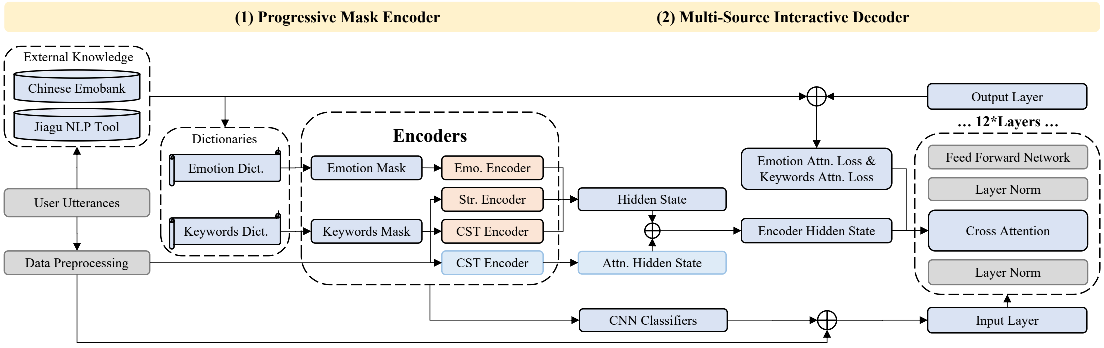

Figure 2: Overall architecture of CSD.

i.e., $\mathcal{U} = [U_1, ..., U_M]$. Following (Lin et al., 2019), we flat $\mathcal{U}$ into a token sequence and insert a CLS token at the start of the token sentence, i.e., $\mathcal{U} = [\mathrm{CLS}, \mathrm{x}_1, ..., \mathrm{x_m}]$.

$$\mathrm{h_i} = \mathrm{LN}(\mathrm{x_i^{l-1}} + \mathrm{MHAtt}(\mathrm{x_i^{l-1}})) \tag{1}$$

$$\widetilde{\mathrm{x}}_{\mathrm{i}}^{\mathrm{l}} = \mathrm{LN}(\mathrm{h_i} + \mathrm{FFN}(\mathrm{h_i})) \tag{2}$$

where LN is the layer normalization proposed by (Ba et al., 2016). MHAtt is multi-head attention, which runs through an attention mechanism several times in parallel (Vaswani et al., 2017). FFN is a two-layer feed-forward network with ReLU as the hidden activation function. The encoder contains l layers. $\mathrm{h_i}$ is the hidden state of the i-th token and $\widetilde{\mathrm{x}}_{\mathrm{i}}^{\mathrm{l}}$ is the embedding with context of the i-th token at the l layer. The obtained context representations are denoted as $\mathbf{C_u} = [\widetilde{\mathrm{x}}_0, ..., \widetilde{\mathrm{x}}_\mathrm{m}]$. Let $\mathrm{l_{cs}}$ be the label of the therapy classification result, i.e.,

$$\mathrm{l_{cs}} = \mathrm{CNN}(\mathbf{C_u}) \tag{3}$$

where CNN is a TextCNN classifier (Kim, 2014) with convolution kernel sizes (2,3,4) and 256 convolution kernels. Similarly, $\mathrm{l_{emo}}$ and $\mathrm{l_{str}}$ are obtained, representing the labels of the emotion classification result and the strategy classification result, respectively.

3.3 Multi-Source Interactive Decoder

In the decoder generation module, we further insert a SEP token at the end of every utterance in order to distinguish the utterances between SPEAKER and LISTENER in multiple rounds of conversation, i.e., $\mathcal{U} = [\mathrm{CLS}, \mathrm{x}_1, ..., \mathrm{x_m}, \mathrm{SEP}]$.

In order to generate responses more suitable for our scenario, encoders, external knowledge and decoder interact in three aspects: (1) input layer; (2) cross-attention mechanism; (3) attention loss.

Input Layer. We take the therapy label $\mathrm{l_{cs}}$, emotional label $\mathrm{l_{emo}}$, and strategy label $\mathrm{l_{str}}$ that encoder classification models generate as three tokens (t_{emo}, t_{cs}, t_{str}) and append them at the end of each utterance. We can then obtain decoder input tokens $\mathcal{Y} = [y_1, ..., y_j, t_{emo}, t_{cs}, t_{str}]$. To represent sentences and knowledge, we first use a word embedding layer, a positional embedding layer to convert each token into vectors (Vaswani et al., 2017), i.e., $\mathbf{E}_W(y_j) \in \mathbb{R}^d$, $\mathbf{E}_P(y_j) \in \mathbb{R}^d$, where d is the dimensionality of embeddings. $\mathbf{y}_j$ is computed as follows: $[y_1, ..., y_j, t_{emo}, t_{cs}, t_{str}]$ is the composition of two types of embeddings.

Cross-Attention Mechanism. We first train an extra encoder that flattens the input data (the format of the data is the same as that of the decoder input), and get the corresponding hidden states $\mathrm{he_j}$:

$$\mathrm{he_j} = \mathrm{LN}(\mathrm{y_j^{l-1}} + \mathrm{MHAtt}(\mathrm{y_j^{l-1}})) \tag{4}$$

In order to more reasonably embed the representation of SPEAKER/LISTENR's utterances generated by encoders into the decoder through cross-attention mechanism, we extract the hidden states from the encoder classification models to replace the hidden states of the labels position ($\mathrm{he_{emo}}$, $\mathrm{he_{cs}}$, $\mathrm{he_{str}}$) generated by extra encoder, forming new encoder hidden states embedded in the cross attention of decoder.

Attention Loss. Since humans naturally pay extra attention to emotional support and therapy information during a conversation, we enforce an emotional attention loss and keyword attention loss in order to focus on those words with higher emotion intensity values and keyword intensity values. Emotional intensity values and keyword intensity values are obtained from Chinese Emobank and Jiagu, respectively.

To highlight emotional information, we compute emotion intensity values for dialogue words and external concepts y_j:

$$\eta_{emo}(y_j) = \frac{(V_a(y_j) + A_r(y_j)) - 2 * \mathrm{R_{min}}}{\mathrm{R_{max}} - \mathrm{R_{min}}} \quad (5)$$

where $V_a(y_j)$ and $A_r(y_j)$ denote the mean values of valence and arousal dimensions of word y_j, respectively. $\mathrm{R_{min}}$ and $\mathrm{R_{max}}$ represent the minimal and maximal values of the value range, respectively. If y_j is not in Chinese EmoBank, $\eta_{emo}(y_j)$ will be set to 0.

To highlight keyword information, keyword intensity values for dialogue words y_j are used based on Jiagu's keyword extraction function:

$$\eta_{kw}(y_j) = \mathrm{softmax(y_j)} \quad (6)$$

where the softmax operation calculates a probability for every word and the probabilities of all the words add up to one.

Emotion loss $\mathcal{L}_{emo}$ and keywords loss $\mathcal{L}_{kw}$ are calculated using Mean Square Error (MSE).

$$\mathcal{L}_{emo} = \frac{1}{e} \times \sum_{i=1}^{e} (\eta_{emo}(y_j) - a_j)^2 \quad (7)$$

$$\mathcal{L}_{kw} = \frac{1}{e} \times \sum_{i=1}^{e} (\eta_{kw}(y_j) - a_j)^2 \quad (8)$$

where a_j is the attention weight of each word in the utterance calculated by the attention output tensors.

When the model generates the response, we use a sampling method to generate the next j-th token Given $\mathcal{U}$ and tokens t_{emo}, t_{cs} and t_{str}, our multi-source interactive decoder aims to generate a n-length response $\mathcal{Y} = \{y_1, ..., y_n\}$ through maximizing the probability $\mathrm{P}(\mathcal{Y}|\mathcal{U}, \mathrm{t_{emo}}, \mathrm{t_{cs}}, \mathrm{t_{str}}) = \prod_{\mathrm{n=1}}^{\mathrm{N}} \mathrm{P}(\mathrm{y_n}|\mathrm{y_{<n}}, \mathcal{U}, \mathrm{t_{emo}}, \mathrm{t_{cs}}, \mathrm{t_{str}})$.

Like most dialogue generation tasks, standard maximum likelihood estimator (MLE) is used as the optimization objective:

$$\mathcal{L}_{gen} = -\mathrm{log}(\mathrm{P}(\mathcal{Y}|\mathcal{U}, \mathrm{t_{emo}}, \mathrm{t_{cs}}, \mathrm{t_{str}})) \quad (9)$$

Eventually, a joint loss function is defined to jointly minimize the emotion attention loss (Eq. 7), the keywords attention loss (Eq. 8) and the generation loss (Eq. 9) as follows:

$$\mathcal{L} = \gamma_1 * \mathcal{L}_{gen} + \gamma_2 * \mathcal{L}_{emo} + \gamma_3 * \mathcal{L}_{kw} \quad (10)$$

where γ_1, γ_2 and γ_3 are hyper-parameters.

3.4 Training

We divide training into three phases as follows: (1) Encoders are used for context training based on the progressive mask method. Two pre-trained encoder models are trained based on sentiment dictionary and keyword dictionary, respectively. (2) Therapy classification and strategy classification tasks are realized on the basis of the encoder trained according to the keyword dictionary. The task of emotion classification is realized based on the encoder trained according to the emotion dictionary. (3) We use the flatten data as the training data of the encoder, making the batch size and input data consistent with the decoder. Then the hidden state of the last layer of the encoder is interacted with the decoder through the cross attention mechanism.

4 Experiments

4.1 Implementation Details

We conduct experiments on the CSConv dataset. For the encoder module of the CSD, the pre-trained model is bert-base-chinese[4], and the decoder module is gpt2-chinese-cluecorpussmall (**?**). Most of the hyperparameters are the same as those in decoder chitchat[5]. In the progressive mask encoder trained based on the keyword dictionary, the ratios of masked entities and sentences (i.e., λ_1, λ_2, λ_3, λ_4 and λ_5) are set as 0.9, 0.9, 0.9, 0.9 and 0.4, respectively. Based on the emotion dictionary, λ_1, λ_2, λ_3, λ_4 and λ_5 are set as 0.5, 0.5, 0.4, 0.3 and 0.2, respectively. Loss weights, namely γ_1, γ_2 and γ_3, are set as 1, 0.5 and 0.5, respectively. We implement all models with PyTorch (Paszke et al., 2019) on four NVIDIA A100 GPUs, and train the models using AdamW optimizer (Loshchilov and Hutter, 2017) with a batch size of 4. We vary the learning rate during training following (Vaswani et al., 2017). For inference, we set the temperature as 0.7, top-k as 8 and top-p as 0.5. The training time for the encoder of the CSD is about 2 minutes and that for the decoder is about 33 minutes. In testing different models, we use NLTK packages to compute the Bleu metric and bert-score package to compute BERTScore. We set the smooth function of NLTK to method 7, and the model used in computing the bert-score is bert-base-chinese.

[4] https://huggingface.co/bert-base-chinese

[5] https://github.com/yangjianxin1/GPT2-chitchat

Models	Therapy Accuracy	Emotion Accuracy	Strategy Accuracy
Transformer	83.67	85.10	91.63
BERT	85.71	87.76	94.49
BERT+CNN	84.90	87.35	94.29
CSD	**87.14**	88.37	**94.69**
CSD (+CNN)	85.92	**88.57**	94.08

Table 4: Evaluation results between baselines and the encoder module of our CSD.

Models/Products	Bleu-2	Bleu-4	BERTScore	Distinct-1	Distinct-2	Empathy	Support	Fluency
CDialGPT$_{\text{base}}$	17.55	6.22	57.70	8.61	29.34	3.10	3.11	3.20
CDialGPT$_{\text{large}}$	15.05	5.47	57.81	**9.61**	**32.62**	3.17	3.19	3.17
GPT2-chitchat	34.61	21.04	66.37	5.29	17.85	3.31	3.37	3.33
Distil-cluecorpussmall	39.94	25.30	69.41	6.44	22.47	3.27	3.31	3.29
Cluecorpussmall	41.04	26.59	68.65	6.79	23.75	3.39	3.32	3.39
CSD	**45.53**	**30.90**	**74.61**	6.90	27.04	**3.61**	**3.49**	**3.57**

Table 5: Evaluation results between baselines and our CSD. The first five metrics are automatic metrics, while the last three metrics are human metrics. **Bold face** indicates leading results in terms of the corresponding metric.

4.2 Automatic Evaluation

For encoder classification, to evaluate the model at the emotional level, we adopt **Emotion Accuracy** as the evaluation metric between the ground truth emotion labels and the predicted emotion labels. **Therapy Accuracy** and **Strategy Accuracy** are similar evaluation metrics to emotion accuracy.

For decoder generation, we employ **BLEU** (Papineni et al., 2002), an algorithm for evaluating the text quality, as the metric. Since BLEU cannot perfectly reflect the quality of generated results (Liu et al., 2016), we adopt **BERTScore** (Zhang et al., 2019a) to compare the similarity between embeddings of a generated sentence and the reference sentence. **Distinct-1 / Distinct-2** (Li et al., 2016) is the proportion of the distinct uni / bi-grams in all the generated results, that indicate the diversity.

4.3 Human Evaluation

To qualitatively examine model performance, we also conduct human evaluations. We sample some dialogues from the CSD and the baselines. We find 6 elders and their relatives to evaluate the responses generated by different models. All models are evaluated in terms of **Empathy**, **Support** and **Fluency**. Empathy measures whether LISTENER understands SPEAKER's feelings. Support measures whether LISTENER gives SPEAKER reasonable help and comfort. Fluency measures the grammatical correctness and readability of the SPEAKER's responses. Each metric is rated on five-scale, where 1, 3 and 5 indicate unacceptable, moderate and excellent performance, respectively.

4.4 Baselines for Comparison

We conduct extensive experiments to compare the encoder module of the CSD against the following representative baselines: (1) **Transformer** (Vaswani et al., 2017): A transformer-based encoder-decoder model. (2) **BERT** (Kenton and Toutanova, 2019): BERT is a context-aware encoder, and is good at processing downstream tasks, like classification. (3) **BERT+CNN**[6]: The model is the embedding with contextual meaning output by BERT, which is input into a CNN classifier for classification.

We conduct extensive experiments to compare the decoder generation module of CSD against the following representative baselines: (1) **CDialGPT-base** (Wang et al., 2020a): The model is a 12-layer GPT model trained on the LCCC-base dataset. (2) **CDialGPT-large** (Wang et al., 2020a): The model is a 12-layer GPT model trained on the LCCC-large dataset. (3) **GPT2-chitchat**[7]: The model is a 10-layer GPT-2 trained on 500,000 chitchat corpus. (4) **Distil-cluecorpussmall** (Radford et al., 2019; **?**): The model is a 6-layer GPT-2 trained on the CLUECorpusSmall (Xu et al., 2020) corpus. (5) **Cluecorpussmall** (Radford et al., 2019; **?**): The model is a 12-layer GPT-2 trained on the CLUECorpusSmall corpus.

To better analyze the influence of different components in the CSD, we also conduct an ablation

[6] https://github.com/649453932/Bert-Chinese-Text-Classification-Pytorch

[7] https://github.com/yangjianxin1/GPT2-chitchat

study as follows: (1) **w/o** NM: The CSD model uses only traditional BERT instead of BERT trained using the progressive mask method. (2) **w/o** IL: The CSD model only splices three classification result labels after utterance as the train data. (3) **w/o** CA: The CSD model only interacts with encoder through the cross-attention mechanism. (4) **w/o** AL: The CSD model only adds the attention loss to embed external knowledge.

Models	Bleu-2	Bleu-4	BERTScore	Distinct-2
CSD	**45.53**	**30.90**	**74.61**	27.04
w/o NM	44.75	30.42	74.27	26.77
w/o IL	42.88	30.53	73.22	22.71
w/o CA	43.39	28.73	72.79	**29.54**
w/o AL	43.66	28.91	70.97	23.20

Table 6: Ablation test of different components.

Models	Win	Loss	Tie
CSD vs CDialGPT$_{\text{base}}$	69.0	20.7	10.3
CSD vs CDialGPT$_{\text{large}}$	65.5	20.7	13.8
CSD vs GPT2-chitchat	55.2	17.2	27.6
CSD vs Distil-cluecorpussmall	48.3	27.6	24.1
CSD vs Cluecorpussmall	41.4	31.0	27.6

Table 7: Result of human A/B test.

4.5 Experimental Results and Analysis

Automatic evaluations. In Table 4, we observe that the encoder module of the CSD is better than the other baselines in therapy, emotion, support strategy recognition accuracy. In Table 5, we observe that the CSD outperforms strong baselines in terms of Bleu and BERTScore. Because CSD models extensive therapy principle and emotional support strategy and there is less language diversity associated with therapy principle and emotional support strategy, the diversity of response is weaker than that of CDialGPT$_{\text{base}}$ and CDialGPT$_{\text{large}}$.

We also perform an ablation study for better understanding the contributions of the main modules of the CSD model. As shown in Table 6, CSD outperforms all other models (w/o NM, w/o IL, w/o CA, w/o AL) in Bleu and BERTScore. However, due to therapy principle and emotional support strategy intervening in the generation of decoders, the diversity of response generation decreases. Only the case of w/o CA model involving a small number of therapies and support strategies achieves high diversity of generated responses.

Human evaluation. Table 5 illustrates that CSD obtains the best performance on Empathy, Support and Fluency scores. Additionally, we carry out pairwise response comparison to directly compare the dialogue quality gains in Table 7. The results confirm that the responses from CSD are more preferred by human judges.

4.6 External Knowledge Analysis

We introduce external knowledge in three ways: training encoders by using external knowledge to progressively mask entities and sentences (w/o NM), intervening GPT-2 generation by classification labels (w/o IL), and paying more attention to emotional words and keywords by calculating the weight of words (w/o AL). To further investigate the impact of introduced knowledge, we test different components of CSD as shown in Table 6. However, the distinct metrics of these models are lower than models without embedded knowledge (w/o CA). Because w/o NM has more knowledge embedded than w/o IL and w/o AL and distinct metric of w/o NM is also significantly improved compared with w/o IL and w/o AL, it concluded that the generated response diversity decreases when little external knowledge is embedded, but with the increase of embedded knowledge, diversity of the generated response also increases.

4.7 Case Study

For decoder generation evaluation, Table 8 shows two examples generated by CSD and other baselines. In the first case, CSD generates an informative response with proper therapy principle and emotional support, which stimulates thinking of the elder through implicit empathy and further questioning. However, baselines with only the decoder part fail to express responses with the therapy principle and emotional support. In the second case, CSD generates a response with continuous questions, which further stimulates thinking of elder. Both cases show that CSD can generate responses with therapy principle and emotional support.

5 Related Work

5.1 Cognitive Training Dialogue System

With the increasing popularity of NLP, dialogue systems have progressed from exploiting simple neural networks (Lee et al., 2016) to large-scale pretrained models (Vlasov et al., 2019; Zhang et al., 2019b; Ni et al., 2022). Currently, while English dialogue systems dominate, there also exist Chi-

History and Labels	X_1: Where did you get your hair cut? (Inquiry, None, Question) X_2: At the community center. (Expression, None, None) X_3: Wow, is there anyone at the community center who cuts hair? (Inquiry, Surprise, Question) X_4: Yes, it's very cheap, five dollars. (Expression, None, None)
Gold	**Wow, that's cheap. Who cut your hair?**
$\text{CDialGPT}_{\text{base}}$	Well expensive!
$\text{CDialGPT}_{\text{large}}$	Yes, it's very nice.
GPT2-chitchat	Yes, it's great!
Distil-cluecorpussmall	How do you cut your hair?
Cluecorpussmall	I feel the same way.
CSD	**Wow, five dollars, that means they are actually volunteering, right?**
History and Labels	X_1: I've just heard "Nanping Evening Bells". (Expression, None, Self-disclosure)
Gold	**This song seems very old. Have you heard this song before?**
$\text{CDialGPT}_{\text{base}}$	I just listened to it.
$\text{CDialGPT}_{\text{large}}$	Ha ha, I just heard that too.
GPT2-chitchat	Have you ever heard the song "Nanping Evening Bells"?
Distil-cluecorpussmall	Have you heard this song?
Cluecorpussmall	Do you seem to have heard this song before?
CSD	**Who sings this song? Why is it called "Nanping Evening Bells"?**

Table 8: Generated responses (translated from Chinese to English) from CSD and baseline models. The emotion classification, therapy principle, support strategy are labeled in the parentheses after the utterances.

nese ones[8] (Wang et al., 2020b; Zhou et al., 2021; Gu et al., 2022). Most of these dialogue systems are for ordinary people, and there are few cognitive recovery dialogue systems for elders with cognitive impairment. Most of the existing dialogue systems for elders focus on specific functions, such as storytelling (Tokunaga et al., 2019, 2021), robotic dialogue based on photos (Tokunaga et al., 2021), etc. There are also dialogue systems for Metamemory therapy (Kim et al., 2021b). Few dialogue systems exist on cognitive stimulation (Navarro et al., 2018), let alone in Chinese.

5.2 Empathetic Dialogue, Emotional Support Conversation and Related Datasets

With the rise of data-driven learning methods (Vaswani et al., 2017), there are more and more studies on open-domain dialogue generation patterns (Dinan et al., 2018; Kann et al., 2022). In order to generate an emotional response, many methods automatically recognize the current user's emotional state through the conversation (Sabour et al., 2022; Gao et al., 2021; Kim et al., 2021a; Shen et al., 2021; Welivita and Pu, 2020; Lin et al., 2019). (Li et al., 2020) propose a multi-resolution adversarial framework which considers multi-granularity emotion factors and user feedback. (Li et al., 2022) propose a knowledge-aware empathetic dialogue generation method, which interferes with generation by embedding external knowledge into the Transformer model via diagrams. Some studies (Sharma et al., 2020, 2021) on empathetic dialogue technologies have also been applied to mental health. About dataset, EMPATHETICDIALOGUES (Rashkin et al., 2019) is the benchmark of the empathetic dialogue datasets, but there exist very few relevant datasets in Chinese.

Different from empathetic dialogue, emotional support conversation can provide emotional support and problem solving in addition to empathetic responses (Liu et al., 2021). Because the field is new, there are a few studies on emotional support conversation (Tu et al., 2022; Peng et al., 2022; Xu et al., 2022). (Tu et al., 2022) propose MISC, which is a mixed strategy-aware model integrating COMET for emotional support conversation. ESConv (Liu et al., 2021) is the benchmark of the emotional support conversation datasets, but there is no Chinese emotional support conversation dataset.

6 Conclusion and Outlook

In this paper, we construct a Chinese CS conversation dataset and propose a multi-source knowledge fusion method for CS dialogue. Experimental results show that CSD outperforms state-of-the-art models in terms of both automatic and human evaluations. Extensive experiments verify the effectiveness of the progressive mask method and the three interaction ways of multi-source interactive decoder in CSD. As for future work, we plan to construct larger datasets of Mandarin and Cantonese

[8] https://github.com/yangjianxin1/GPT2-chitchat

CS conversations to train models, and address the issue of therapy principle, emotional support recognition in reference context in dialogue.

Limitations

The current dialogue system is mainly based on deep neural network, like transformer structure, which often requires a large number of data sets for training model. However, there are still some deficiencies in our dataset. We will further label and create more dataset to train model. In addition, in order to improve the quality of dialogue, our model parameters are relatively large, which affect the speed of dialogue generation to some extent. We will explore some methods, such as knowledge distillation, to reduce model parameters to improve the speed of dialogue generation on the premise of keeping the quality of dialogue generation unchanged.

Ethics Statement

We have sought to ethically conduct this study, including transparently communicating with data annotators about data use and study intent, and finding suitable elders to conduct human tests of the dialogue systems, compensating workers and elders at a reasonable hourly wage. We have obtained study approval from the ethics review board.

Acknowledgements

We want to thank our anonymous AC and reviewers for their feedback. This work was supported in part by grants from Hong Kong Research Grants Council (RGC) under the contracts HKU 17203522 and 17207621.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Thaís Cristina Galdino De Oliveira, Fernanda Cabral Soares, Liliane Dias E Dias De Macedo, Domingos Luiz Wanderley Picanço Diniz, Natáli Valim Oliver Bento-Torres, and Cristovam Wanderley Picanço-Diniz. 2014. Beneficial effects of multisensory and cognitive stimulation on age-related cognitive decline in long-term-care institutions. *Clinical Interventions in Aging*, pages 309–321.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *CoRR*, abs/1811.01241.

Lea Ellwardt, Marja Aartsen, Dorly Deeg, and Nardi Steverink. 2013. Does loneliness mediate the relation between social support and cognitive functioning in later life? *Social science & medicine*, 98:116–124.

Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 807–819.

Linda J Garcia. 2022. The usefulness of useless conversation: An avenue for connection and autonomy for older adults. In *Well-being In Later Life*, pages 53–64. Routledge.

Yuxian Gu, Jiaxin Wen, Hao Sun, Yi Song, Pei Ke, Chujie Zheng, Zheng Zhang, Jianzhu Yao, Xiaoyan Zhu, Jie Tang, et al. 2022. Eva2. 0: Investigating open-domain chinese dialogue systems with large-scale pre-training. *arXiv preprint arXiv:2203.09313*.

Katharina Kann, Abteen Ebrahimi, Joewie Koh, Shiran Dudy, and Alessandro Roncone. 2022. Open-domain dialogue generation: What we can do, cannot do, and should do next. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 148–165.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, pages 4171–4186.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021a. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. *EMNLP*.

Jeongsim Kim, EunJi Shin, KyungHwa Han, Soowon Park, Jung Hae Youn, Guixiang Jin, Jun-Young Lee, et al. 2021b. Efficacy of smart speaker–based metamemory training in older adults: Case-control cohort study. *Journal of medical Internet research*, 23(2):e20177.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Hanbit Lee, Yeonchan Ahn, Haejun Lee, Seungdo Ha, and Sang-goo Lee. 2016. Quote recommendation in dialogue using deep neural network. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 957–960.

Lung-Hao Lee, Jian-Hong Li, and Liang-Chih Yu. 2022. Chinese emobank: Building valence-arousal resources for dimensional sentiment analysis. *Transactions on Asian and Low-Resource Language Information Processing*, 21(4):1–18.

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.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. Empdg: Multiresolution interactive empathetic dialogue generation. *COLING*.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. *36th Association for the Advancement of Artificial Intelligence*.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. *CoRR*, abs/1908.07687.

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.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. pages 3469–3483.

Yingxu Liu, Shu Zhang, Yasutake Tomata, Tatsui Otsuka, Dieta Nurrika, Yumi Sugawara, and Ichiro Tsuji. 2020. Emotional support (giving or receiving) and risk of incident dementia: The ohsaki cohort 2006 study. *Archives of Gerontology and Geriatrics*, 86:103964.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Javier Navarro, Faiyaz Doctor, Víctor Zamudio, Rahat Iqbal, Arun Kumar Sangaiah, and Carlos Lino. 2018. Fuzzy adaptive cognitive stimulation therapy generation for alzheimer's sufferers: Towards a pervasive dementia care monitoring platform. *Future Generation Computer Systems*, 88:479–490.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2022. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial Intelligence Review*, pages 1–101.

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.

Jeong-Mo Park, Mi-Won Kim, and Hee-Young Shim. 2019. Effects of a multicomponent cognitive stimulation program on cognitive function improvement among elderly women. *Asian Nursing Research*, 13(5):306–312.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Wei Peng, Ziyuan Qin, Yue Hu, Yuqiang Xie, and Yunpeng Li. 2022. Fado: Feedback-aware double controlling network for emotional support conversation. *arXiv preprint arXiv:2211.00250*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.

Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*, WWW '21, page 194–205, New York, NY, USA. Association for Computing Machinery.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Lei Shen, Jinchao Zhang, Jiao Ou, Xiaofang Zhao, and Jie Zhou. 2021. Constructing emotion consensus and utilizing unpaired data for empathetic dialogue generation. *EMNLP*.

Seiki Tokunaga, Katie Seaborn, Kazuhiro Tamura, and Mihoko Otake-Matsuura. 2019. Cognitive training for older adults with a dialogue-based, robot-facilitated storytelling system. In *International Conference on Interactive Digital Storytelling*, pages 405–409. Springer.

Seiki Tokunaga, Kazuhiro Tamura, and Mihoko Otake-Matsuura. 2021. A dialogue-based system with photo and storytelling for older adults: toward daily cognitive training. *Frontiers in Robotics and AI*, page 179.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. Misc: A mixed strategy-aware model integrating comet for emotional support conversation. *60th Annual Meeting of the Association for Computational Linguistics*.

Helma van Rijn, Joost van Hoof, and Pieter Jan Stappers. 2010. Designing leisure products for people with dementia: Developing "the chitchatters"game. *American Journal of Alzheimer's Disease & Other Dementias®*, 25(1):74–89.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vladimir Vlasov, Johannes E. M. Mosig, and Alan Nichol. 2019. Dialogue transformers. *CoRR*, abs/1910.00486.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020a. A large-scale chinese short-text conversation dataset. In *NLPCC*.

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

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. *CoRR*, abs/2012.04080.

Liang Xu, Xuanwei Zhang, and Qianqian Dong. 2020. Cluecorpus2020: A large-scale chinese corpus for pre-training language model. *ArXiv*, abs/2003.01355.

Xiaohan Xu, Xuying Meng, and Yequan Wang. 2022. Poke: Prior knowledge enhanced emotional support conversation with latent variable. *arXiv preprint arXiv:2210.12640*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019a. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019b. Dialogpt: Large-scale generative pre-training for conversational response generation. *CoRR*, abs/1911.00536.

Hao Zhou, Pei Ke, Zheng Zhang, Yuxian Gu, Yinhe Zheng, Chujie Zheng, Yida Wang, Chen Henry Wu, Hao Sun, Xiaocong Yang, et al. 2021. Eva: An open-domain chinese dialogue system with large-scale generative pre-training. *arXiv preprint arXiv:2108.01547*.