Using Multi-Encoder Fusion Strategies to Improve Personalized Response Selection

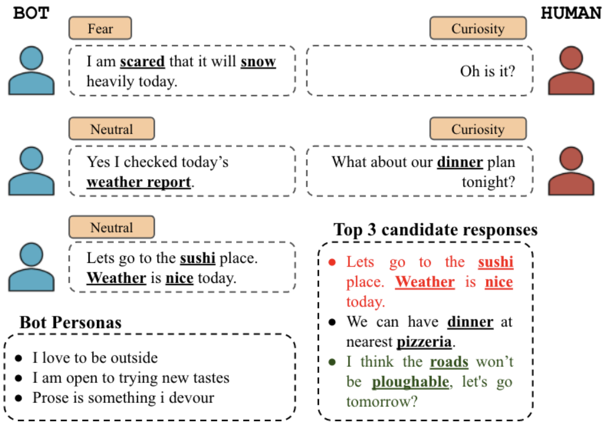
Anonymous ACL submission

Abstract

Personalized response selection systems are generally grounded on *persona*. However, there exists a co-relation between persona and empathy which is not explored well in these systems. Also, *faithfulness* to the conversation context plunges when a contradictory or an off-topic response is selected. This paper makes an attempt to address these issues by proposing a suite of fusion strategies that captures the interaction between persona, emotion, and entailment information of the utterances. Ablation studies were done on `Persona-Chat` dataset show that incorporating emotion, entailment improves the accuracy of response selection. We combine our fusion strategies and concept-flow encoding to train a BERT based model which outperforms the previous methods by margins larger than 2.3% on original personas and 1.9% on revised personas in terms of **hits@1** (top-1 accuracy), achieving a new state-of-the-art performance on the `Persona-Chat` dataset.

1 Introduction

Currently, most response selection systems tend to perform well in most of the cases (Gu et al., 2021a; Zhang et al., 2021b; Gu et al., 2019a, 2020a). On the contrary, these re-ranking systems have poor capability to detect and evade contradictory responses. Often candidate responses directly contradict any of the previous utterances, and any form of contradiction disrupts the flow of conversation. Several efforts have been made to incorporate persona while selecting(Gu et al., 2021b; Zhang et al., 2021a) or generating (Wu et al., 2021) responses. On the other hand, persona is directly correlated with personality (Leary and Allen, 2011), which in turn influences empathy (Richendoller and Weaver III, 1994). (Zhong et al., 2020) presented a multi-domain dataset collected from several empathetic Reddit threads contributing towards persona-based empathetic conversations. However, no work is done to study the emotion-persona interplay in data which is presented in a more natural form. Figure 1 depicts situational emotion can sometime supersede persona to influence response selection. On the contrary, different personality traits are related to emotion regulation difficulties (Pollock et al., 2016). Due to which a person's expected emotion can deviate based on his persona. In addition to that, we also observe concepts that are actively discussed in a conversational flow play an important role, and not much effort is made to incorporate this in response selection.



I am **scared** that it will **snow** heavily today.

Oh is it?

Yes I checked today's **weather report**.

What about our **dinner** plan tonight?

Lets go to the **sushi** place. **Weather** is **nice** today.

Bot Personas

- I love to be outside
- I am open to trying new tastes
- Prose is something i devour

Top 3 candidate responses

- Lets go to the **sushi** place. **Weather** is **nice** today.
- We can have **dinner** at nearest **pizzeria**.
- I think the **roads** won't be **ploughable**, let's go tomorrow?

Figure 1: For this conversation the selected candidate response directly contradicts the context. Also, the bot's persona is influencing the response selection, while the situational emotions and concepts gets ignored. The underlines phrases/words denotes the concepts.

Model	Emotion Inappropriate(%)	Contradictory(%)	Off-topic(%)
BERT-CRA	7.35	11.88	12.3

Table 1: Statistics of issues reported in the test split of Persona-Chat inferred by BERT-CRA (Gu et al., 2021b) [1]

[1] Insights drawn from human evaluation done on 500 randomly selected data-points from self-persona original and partner-persona original sets of `Persona-Chat`

The significance of these problem can be inferred from Table 1. So, to increase the usability of the personalized response selection systems, all these fundamental problems need to be addressed. In order to model emotion-persona interaction, context-response entailment, and concept-flow we automatically annotate Persona-Chat (Zhang et al., 2018) data set using a series of classifiers and rule-based modules. To compare the ability of annotated features to enhance the emotion-persona interaction, contradiction avoidance, and to adhere to the concept-flow, we perform preliminary experiments by devising independent encoders based on BERT. Our baseline model extends BERT-CRA (Gu et al., 2021b) where we introduce an additional bot-encoder to better represent the bot-utterances. Subsequently, we propose three fusion strategies, emotion-aware(EmA), entailment-aware(EnA), persona-entailment-aware(P-EmA). These fusion strategies are designed based on emotion-persona interaction or persona-entailment information. Along with these fusion strategies we propose a concept-flow encoding technique that matches relevant concepts from the context and candidate responses.

We test our proposed methods on the Persona-Chat dataset with our automatic annotation. The results show that a model trained on a combination of our proposed fusion strategies outperforms the current state-of-the-art model by a margin of 2.3% in terms of top-1 accuracy **hits@1**.

In summary, the contributions of this paper are three-fold. (1)Automatically annotate Persona-Chat dataset, with utterance level emotion, entailment, and concept information to provide extra supervision. (2) A suite of fusion strategies and a concept-flow encoder which are designed and implemented into a series of models, aiming to explore the impact of emotion, entailment, and concept-flow in the task of response selection. (3) Experimental results demonstrate that our proposed models outperform the existing state-of-the-art models by significant margins on the widely used Persona-Chat response selection benchmark.

2 Related Works

2.1 Personalized Response Selection

Chit-chat models suffer from a lack of a consistent personality as they are typically trained over many dialogues, each with different speakers, and a lack of explicit long-term memory as they are typically trained to produce an utterance given only a very recent dialogue history. (Li et al., 2016) proposed a persona-based neural conversation model to capture individual characteristics such as background information and speaking style. (Zhang et al., 2018) has constructed Persona-Chat dataset to build personalized dialog systems, this is by far the largest public dataset containing million-turn dialog conditioned on persona. Many benchmarks have been established for this dataset, for example, (Mazaré et al., 2018) proposed the fine-tuned Persona-Chat (FT-PC) model which first pre-trained models using a large-scale corpus based on Reddit to extract valuable dialogues conditioned on personas, and then fine-tuned these pre-trained models on the Persona-Chat dataset. (Wolf et al., 2019; Liu et al., 2020) also employed the pre-trained language model(GPT) for building personalized dialogue agents. (Gu et al., 2020c) proposed filtering before iteratively referring (FIRE) to ground the conversation on the given knowledge and then perform the deep and iterative matching. (Gu et al., 2021b) explored a new direction by proposing four persona fusion strategies and thereby incorporating partner persona in response selection.

2.2 Faithfulness to Context

Faithfulness in conversational systems to conversation context or knowledge is a very broad topic that can range from decreasing fact hallucination(Chen et al., 2021), reducing contradictory responses, staying on topic, etc. (Rashkin et al., 2021) has used additional inputs to act as stylistic controls that encourage the model to generate responses that are faithful to a provided evidence or knowledge. However, no one has studied the level of faithfulness the current personalized response selection systems exhibit with respect to the conversation history. Thus, this paper attempts to thoroughly explore the impact of utilizing utterance level emotions, entailment, and concepts on the performance of personalized response selection.

3 Dataset

In this work, we extend Persona-Chat (Zhang et al., 2018) and augment it with a series of annotators. The dataset consists of 8939 complete dialogues for training, 1000 for validation, and 968 for testing. Responses are selected at every turn of a conversation sequence, which results in 65719 context-responses pairs for training, 7801 for validation, and 7512 for testing in total. The positive and negative responses ratio is 1:19 in the training, validation, and testing sets. There are 955 possible personas for training, 100 for validation, and 100 for testing, each consisting of 3 to 5 profile sentences. To make this task more challenging, a revised version of persona descriptions is also provided by rephrasing, generalizing, or specializing the original ones.

4 Automatic Dataset Annotation

We have annotated the `Persona-Chat` with the help of a series of automatic annotation schemes. Since we are studying the effect of emotions in personalized response selection, we assign emotion labels to the personas, context-utterances, and candidate responses using an emotion classifier. To incorporate the entailment information while selecting responses, personas and utterances were annotated using an entailment classifier. Finally, to match meaningful concepts appearing in the context and response we follow a multi-layer keyword mining strategy.

4.1 Emotion

We trained an emotion classifier on `GoEmotions` dataset (Demszky et al., 2020). This dataset contains 58k English Reddit comments, labeled for 27 emotion categories or Neutral. We fine-tuned `RoBERTa` using this dataset. We saved the checkpoint with the best Macro F1 of 49.4% and used this for annotating each utterance. Since the performance of emotion classifier is not that significant, we only consider the labels which can be predicted with more than 90% confidence.

4.2 Entailment

For entailment annotation, we have used an ensemble of two models. The first one is an off-the-self `RoBERTa` based model trained on Stanford Natural Language Inference (SNLI) corpus (MacCartney and Manning, 2008) released by AllenAI[2]. Second model is also a `RoBERTa` based model fine-tuned on a recently released NLI dataset, DECODE (Nie et al., 2020). During inference, we take a weighted average of both the probabilities from the two models. The second model is given a higher preference with 80% weightage to its probabilities. The entailment label is assigned to every persona-response and utterance-response pair.

4.3 Concept Mining

We mine keywords and key phrases from the persona sentences, utterances, and responses denoted as $\{pk_i\}_{i=1}^{N_{pk}}$, $\{uk_i\}_{i=1}^{N_{uk}}$, $\{rk_i\}_{i=1}^{N_{rk}}$ respectively. We follow the techniques proposed in (Tang et al., 2019) to extract the first level of keywords. Subsequently, we expand the concepts lists by extracting key phrases using the RAKE (Rose et al., 2010). We hypothesize that concepts appearing in responses should be adhering to the speaker's persona. So, we prune some of the response/ context keywords by calculating the average of Point-wise Mutual Information score between persona keywords and response/ context keywords $\sum_{j=1}^{N_{pk}} PMI(pk_j, rk_i)/N_{pk}$ and rejecting the concepts which are below a threshold value(λ). Similarly, for response/ concept key-phrases extracted using RAKE, we keep only keep top N key-phrases. Finally, for we combine the persona keywords and context keywords and treat them as context keywords.

5 Methodology

5.1 Problem Definition

Given a data-set $D = \{(C_i, uc_i, p_i, r_i, rc_i, y_i)\}_{i=1}^{N}$ is a set of N tuples consisting context C_i, the persona of the speaker or the partner p_i, response to the context r_i, and the ground truth y_i. Set of concepts appearing in context and a response is denoted by uc_i and rc_i respectively. The context can be represented as $C_i = \{(U_j, Emo_j, Entail_j)\}_{j=1}^{L}$ where U_j is an utterance, Emo_j is the dominant emotion present in U_j and $Entail_j$ is the entailment label of U_j with respect to r_i and. The j^{th} utterance U_j is denoted by $U_j = \{u_1 j, u_2 j, ..., u_M j\}$ which consists of M tokens. Each response r_i contains single utterance, $y_i \in \{0, 1\}$, $Emo_j \in \{0, 1, ...P\}$, and $Entail_j \in$ {`entailment`,`neutral`,`contradiction`} where P are the total number of emotion types possible in the D. The task is to train a matching model for D, $g(C, uc, p, rc, r)$. Given a triple of context-persona-response the goal of the matching model $g(C, uc, p, rc, r)$ is to calculate the degree of match between (C, uc, p) and (rc, r).

[2] https://github.com/allenai/allennlp-models

5.2 Bot Context Encoding

When two users are communicating with each other, often many topics are discussed in parallel and sometimes many utterances might not be relevant for response selection. To account for the model to be aware of the speaker change information, (Gu et al., 2020b) introduced a speaker disentanglement strategy in form of *speaker embeddings* fused with the original token embeddings. Though this technique has proven to improve response selection performance (Gu et al., 2020b; Su et al., 2021), however, the problem of maximum length of positional embeddings still exists. To circumvent this, we have created bot-context encoding, which captures the representation of the bot's turns in the context while ignoring the user's turns. The assumption here is, the bot's turns will be most useful in selecting the relevant response. The input sequence that is sent to BERT to encode bot context is composed as follows:

$$x_{si} = [CLS]u_2[EOU]u_4[EOU]...u_{n-1}[EOU][SEP]r_i[EOU] \quad (1)$$

Where $u_1, u_3, ...u_n$ are bot's utterances in the context, $[EOU]$ is a special token denoting the end of utterance.

The resultant tokens x_{si} are passed through `bert-base-uncased`, the last hidden states of k layers i.e. $\{\boldsymbol{h}^{(l)}_{\boldsymbol{s1}}, \boldsymbol{h}^{(l)}_{\boldsymbol{s2}}, ..\boldsymbol{h}^{(l)}_{\boldsymbol{sT}}\}$, for $l = 1, 2, ..k$ are used in downstream tasks.

5.3 Fusion Strategies

To model the inter-dependencies of the persona, emotion and entailment information we use several fusion strategies. We use BERT (Devlin et al., 2019) as our base sentence encoder. Similar to the Bi-encoder (Humeau et al., 2020) we concatenate context utterances as a single context sentence before passing it into BERT.

5.3.1 Baseline Pipeline

For the baseline, we have extended `BERT-CRA` (Gu et al., 2021b) where persona and context are concatenated to form sequence A and response form sequence B. Then these two sequences are concatenated using $[SEP]$ token. We made two changes to this model, firstly, we have added speaker embeddings along with the original token representation. Secondly, we fuse bot-context encoding as described in the previous section with `BERT-CRA` encoding by doing multi-headed attention between the hidden representation of last k layers of both encoder. Token arrangement is as follows:

$$\begin{aligned} x_{CRAi} \quad = \quad & [CLS]p_1p_2...p_i[EOP]u_1[EOU] \\ & ...u_i[EOU][SEP]r_i[EOU] \qquad (2) \end{aligned}$$

Where $p_1p_2...p_i$ are the personalities of the speaker, $[EOP]$ token denotes end of personality representation, $u_1, u_2, ..u_i$ are the utterances in the context. The resultant tokens x_{CRAi} are passed through `bert-base-uncased`, the hidden states of last k layers i.e. $\{\boldsymbol{h}^{(l)}_{\boldsymbol{c1}}, \boldsymbol{h}^{(l)}_{\boldsymbol{c2}}, ..\boldsymbol{h}^{(l)}_{\boldsymbol{cT}}\}$, for $l = 1, 2, ..k$ are used in downstream tasks.

Interaction Layer : Since we are using a multi-encoder pipeline, it is important to capture the interaction between the encoders. For that, we use multi-head attention between hidden states of speaker context encoder and `BERT-CRA`. For ease of presentation, we denote the whole multi-headed attention layer as $f_{mha}(*, *)$. Then these attention outputs are passed through an aggregation layer which basically concatenates then passes it through a 2 layer feed forward network and finally mean pools across all the layers to get $\boldsymbol{h_d}$. The output is passed through a MLP to get the matching degree with the response.

$$\begin{aligned} \{\widetilde{\boldsymbol{h}}^{(l)}_{\boldsymbol{s1}}, \widetilde{\boldsymbol{h}}^{(l)}_{\boldsymbol{s2}}, ..\widetilde{\boldsymbol{h}}^{(l)}_{\boldsymbol{sT}}\} \quad &= \quad f_{mha}(\{\boldsymbol{h}^{(l)}_{\boldsymbol{s1}}, \boldsymbol{h}^{(l)}_{\boldsymbol{s2}}, ..\boldsymbol{h}^{(l)}_{\boldsymbol{sT}}\}, \\ & \qquad \{\boldsymbol{h}^{(l)}_{\boldsymbol{c1}}, \boldsymbol{h}^{(l)}_{\boldsymbol{c2}}, ..\boldsymbol{h}^{(l)}_{\boldsymbol{cT}}\}) \qquad (3) \\ \{\widetilde{\boldsymbol{h}}^{(l)}_{\boldsymbol{c1}}, \widetilde{\boldsymbol{h}}^{(l)}_{\boldsymbol{c2}}, ..\widetilde{\boldsymbol{h}}^{(l)}_{\boldsymbol{cT}}\} \quad &= \quad f_{mha}(\{\boldsymbol{h}^{(l)}_{\boldsymbol{c1}}, \boldsymbol{h}^{(l)}_{\boldsymbol{c2}}, ..\boldsymbol{h}^{(l)}_{\boldsymbol{cT}}\}, \\ & \qquad \{\boldsymbol{h}^{(l)}_{\boldsymbol{s1}}, \boldsymbol{h}^{(l)}_{\boldsymbol{s2}}, ..\boldsymbol{h}^{(l)}_{\boldsymbol{sT}}\}) \qquad (4) \\ \boldsymbol{h_d} \quad &= \quad MeanPool(\{FFN([\{\widetilde{\boldsymbol{h}}^{(l)}_{\boldsymbol{s1}}, \widetilde{\boldsymbol{h}}^{(l)}_{\boldsymbol{s2}}, ..\widetilde{\boldsymbol{h}}^{(l)}_{\boldsymbol{sT}}\}; \\ & \qquad \{\widetilde{\boldsymbol{h}}^{(l)}_{\boldsymbol{c1}}, \widetilde{\boldsymbol{h}}^{(l)}_{\boldsymbol{c2}}, ..\widetilde{\boldsymbol{h}}^{(l)}_{\boldsymbol{cT}}\}])\}^{k}_{l=0}) \qquad (5) \end{aligned}$$

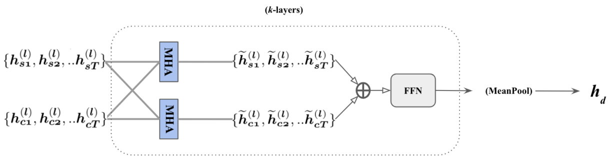

(k-layers)
$\{h^{(l)}_{s1}, h^{(l)}_{s2}, ..h^{(l)}_{sT}\}$
$\{h^{(l)}_{c1}, h^{(l)}_{c2}, ..h^{(l)}_{cT}\}$
MHA
MHA
$\{\widetilde{h}^{(l)}_{s1}, \widetilde{h}^{(l)}_{s2}, ..\widetilde{h}^{(l)}_{sT}\}$
$\{\widetilde{h}^{(l)}_{c1}, \widetilde{h}^{(l)}_{c2}, ..\widetilde{h}^{(l)}_{cT}\}$
$\oplus$
FFN
(MeanPool)
h_d

Figure 2: Interaction Layer

Loss Function: The MLP layer predicts whether a context-persona (C, p) pair matches with the corresponding response r based on the derived features. Subsequently, the output from MLP layer is passed through a softmax output layer to return a probability distribution over all response candidates. All the models described in this paper are learnt using MLP cross-entropy loss. Let Θ be the model parameters then the loss function $\mathcal{L}(D, \Theta)$ for all the models can be formulated as follows:

$$\mathcal{L}(D,\Theta) = -\sum_{(C,p,r,y))\epsilon D} ylog(g(C,p,r)) \tag{6}$$

5.3.2 `BERT-EmA` Emotion Aware Fusion:

In this strategy, an emotion incorporation framework is introduced. Similar to `BERT-CRA` a dual pipeline matching network is followed. The first pipeline encodes the emotional and personality characteristics of both the speaker and listener in the context. While the other encodes the bot-context as described in section 5.2.

To incorporate emotion features in the BERT contextual representation, we attach the most probable emotion tag to each of the utterances. The emotion-infused context representation is then concatenated with the original persona representation like as described in section 5.3.1. The main goal of representing the context in this way is to understand the way the emotions of each utterance interact with the persona of the speaker. The input to emotion encoder is as follows:

$$\begin{aligned} x_{EmAi} \quad = \quad & [CLS]p_1p_2...p_i[EOP][Emo_1]u_1[EOU] \\ & ...[Emo_i]u_i[EOU][SEP]r_i[EOU] \end{aligned} \tag{7}$$

Similar to baseline, the hidden states of last k layers i.e. $\{\boldsymbol{h^{(l)}_{e1}}, \boldsymbol{h^{(l)}_{e2}}, ..\boldsymbol{h^{(l)}_{eT}}\}$, for $l = 1, 2, ..k$ are used in downstream tasks.

5.3.3 `BERT-EnA-P` : Entailment Aware Fusion

In this fusion strategy, the intention is to model the entailment information about each of the utterances and personas with the response. Like `BERT-EmA` we follow a dual encoder pipeline, the first encodes the entailment features and the second encodes the bot-context. To incorporate entailment features into BERT contextual representation, we attach entailment tags i.e. `<contradiction>`, `<entailment>` and `<neutral>` at the start of every utterance and persona. The response is concatenated with the context-entailment representation with a $[SEP]$ token. The input to entailment encoder is as follows:

$$\begin{aligned} x_{EmA-Pi} \quad = \quad & [CLS][Entail_{p1}]p_1...[EOP] \\ & [Entail_1]u_1[EOU] \\ & [Entail_2]u_2[EOU] \\ & ...[Entail_i]u_i[EOU] \\ & [SEP]r_i[EOU] \end{aligned} \tag{8}$$

The hidden states of last k layers i.e. $\{\boldsymbol{h^{(l)}_{en1}}, \boldsymbol{h^{(l)}_{en2}}, ..\boldsymbol{h^{(l)}_{enT}}\}$, for $l = 1, 2, ..k$ are used in downstream tasks.

Finally we experiment with a combined pipeline as depicted in Figure 3.

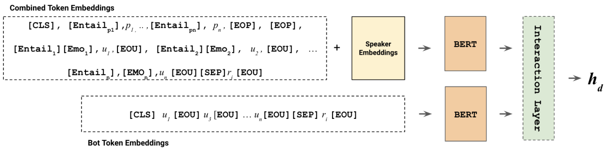

Combined Token Embeddings
[CLS], [Entail_p1], p_1, .., [Entail_pn], p_n, [EOP], [EOP],
[Entail_1][Emo_1], u_1, [EOU], [Entail_2][Emo_2], u_2, [EOU], ...
[Entail_n], [EMO_n], u_n [EOU][SEP] r_i [EOU]
+
Speaker Embeddings
BERT
BERT
Interaction Layer
h_d
[CLS] u_1 [EOU] u_3 [EOU] ... u_n [EOU][SEP] r_i [EOU]
Bot Token Embeddings

Figure 3: Dual encoder pipeline consisting of combination of all the encoding strategies

5.4 Concept-Flow(CF) Interaction

In the earlier section, we describe the process in which we are extracting relevant concepts from the context and the response. Often it is noticed that a relevant response has concepts that are most recently talked about in the context. So, to model that we construct a concept-flow interaction network, where the interaction between the context-concepts and response-concepts are measured and used as a feature in response relevance classification.

Let us consider $\{CC_1, CC_2, ..., CC_n\}$ are concepts extracted from context and $\{RC_1, RC_2, ..., RC_n\}$ are concepts extracted from a response. Now, we pass each of these concepts through a concept encoder f_c to get two sets of concept embeddings $\{\boldsymbol{ec_1}, \boldsymbol{ec_2}, ..., \boldsymbol{ec_n}\}$, , $\boldsymbol{ec_i} \in \mathbb{R}^{d_c}$ and $\{\boldsymbol{rc_1}, \boldsymbol{rc_2}, ..., \boldsymbol{rc_n}\}$, $\boldsymbol{rc_i} \in \mathbb{R}^{d_c}$ for context and response concepts respectively. To learn the context flow representation for each set of concepts, we apply a bi-directional GRU network to capture sequential dependencies between subsequent concepts in a conversational situation. Context-concept and response-concept representation $\boldsymbol{h^{cc}_i}$, $\boldsymbol{h^{rc}_i}$ can be formulated as:

$$\boldsymbol{c_i^{cc}}, \boldsymbol{h_i^{cc}} = \overleftrightarrow{GRU}(\boldsymbol{ec_i}, \boldsymbol{h_{i-1}^{cc}}) \tag{9}$$

$$\boldsymbol{c_i^{rc}}, \boldsymbol{h_i^{rc}} = \overleftrightarrow{GRU}(\boldsymbol{er_i}, \boldsymbol{h_{i-1}^{rc}}) \tag{10}$$

$$\boldsymbol{h_{cc}} = tanh(\sum_{\boldsymbol{j \in 2 * N_l}} \boldsymbol{W_j h_j^{cc}}) \tag{11}$$

$$\boldsymbol{h_{rc}} = tanh(\sum_{\boldsymbol{j \in 2 * N_l}} \boldsymbol{W_j h_j^{rc}}) \tag{12}$$

Where $\boldsymbol{h_i^{cc}} \in \mathbb{R}^{2d_c}$, $\boldsymbol{h_i^{rc}} \in \mathbb{R}^{2d_c}$ are the i - the hidden states and $\boldsymbol{c_i^{cc}} \in \mathbb{R}^{2d_c}$, $\boldsymbol{c_i^{rc}} \in \mathbb{R}^{2d_c}$ are the outputs of the respective GRU encoders, $\boldsymbol{W_j}$ is a learn-able parameter and N_l is the number of layers in each GRUs. To model the interaction between $\boldsymbol{h_i^{cc}}$ and $\boldsymbol{h_i^{rc}}$ we follow the same interaction mechanism described in the earlier section. The output $\boldsymbol{h_{concept}}$ is concatenated with the dual encoder output $\boldsymbol{h_d}$ before passing it through a MLP.

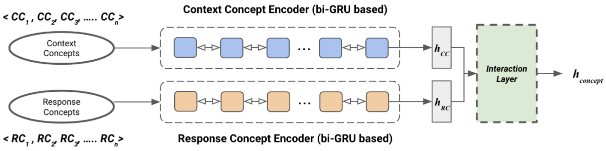

< CC_1 , CC_2, CC_3, ... CC_n>
Context Concept Encoder (bi-GRU based)
Context Concepts
...
h_CC
Interaction Layer
h_concept
Response Concepts
h_RC
< RC_1 , RC_2, RC_3, ... RC_n>
Response Concept Encoder (bi-GRU based)

Figure 4: Concept-flow interaction network, the output of this network $\boldsymbol{h_{concept}}$ can be concatenated with any of the BERT based dual encoder's output($\boldsymbol{h_d}$).

6 Experimental Setup

6.1 Training Details

The ratio of positive to negative samples in the training set is 1:19, so clearly there is a high imbalance in training data. Taking inspirations from (Gu et al., 2021b) we adopted a dynamic negative sampling strategy in which the ratio of positive and negative response is 1:1 in an epoch. For every epoch, we keep the positive response constant and change the negative response, which generates data for 19 epochs. We use `bert-base-uncased` as the base for each of our pretraining-based fusion models. In concept mining strategy we have taken top 3 concepts extracted using RAKE, λ for PMI based scoring was varied from 0.3 to 0.8 with 0.1 step, 0.5 was found optimum. The number of turns in the conversation history used for concept mining varied following this set: $\{2, 3, 4, 5, 6, 7\}$. We preserve the original parameters of `bert-base-uncased`. The number of k-last layers in the interaction layer varied following this set: $\{3, 4, 5, 6\}$, after some initial experimentation 4 was found as the optimum value. The number of heads the multi-head attention layer was kept 8. We use 6-layered version MiniLM(Wang et al., 2020) to encode the concepts, the embedding dimension was 384. The number of layers in the bi-directional GRUs in the concept encoder is 2. A dropout with a rate of 0.7 is applied to the concept encoder hidden representation before we sent it to the interaction layer. AdamW(Loshchilov and Hutter, 2019) optimizer was used for optimization. The initial learning rate was set to 2e-5 and linearly decayed by L2 weight decay. The maximum sequence length was set to 320. The training batch size was 12. The relevance prediction head used a single feed-forward layer with sigmoid activation. All code was implemented in the PyTorch framework. Also, we used 2 NVIDIA RTX A5000 GPUs to train the models. Average training time for 1 epoch was 46 minutes using all our fusion strategies and concept encoding.

6.2 Evaluation Metrics

To ensure results are comparable, we used the same evaluation metrics as in the previous work. Each model aimed to select the best-matched response from available candidates for the given context and persona. We calculated the recall of the true positive replies, denoted as **hits@1**. In addition, the mean reciprocal rank **(MRR)** was also adopted to take the rank of the correct response overall candidates into consideration.

6.3 Comparison Methods

For comparison, we have only selected pretraining-based models only.

- **FT-PC (Mazaré et al., 2018):** employed the "pretrain and fine-tune" framework by first pretraining on a domain-specific corpus, dialogues of which were extracted from Reddit, and then fine-tuning on the Persona-Chat.
- **TransferTransfo (Wolf et al., 2019)**: the paper fine-tunes a transformer model(GPT) using `Persona-Chat` dataset on a multi-task objective which combines several unsupervised task.
- P^2 **Bot (Liu et al., 2020):** incorporates mutual persona to increase quality of dialog generation. It was also initialized and pretrained using GPT on `Persona-Chat` dataset.
- **`BERT-CRA` (Gu et al., 2021b):** This work presents four context-aware persona fusion strategies and the models are initialized and

Model	Self Persona				Partner Persona			
	Original		Revised		Original		Revised	
	hits@1	**MRR**	**hits@1**	**MRR**	**hits@1**	**MRR**	**hits@1**	**MRR**
FT-PC (Mazaré et al., 2018)	-	-	60.7	-	-	-	-	-
DIM (Gu et al., 2019b)	78.8	86.7	70.7	81.2	64.0	76.1	63.9	76.0
TransferTransfo (Wolf et al., 2019)	80.7	-	-	-	-	-	-	-
P2 Bot (Liu et al., 2020)	81.9	-	68.6	-	-	-	-	-
FIRE (Gu et al., 2020c)	81.6	-	74.8	-	-	-	-	-
BERT-CRA (Gu et al., 2021b)	84.3	90.3	79.4	86.9	71.2	80.9	71.8	81.5
BERT-EmA	84.6	90.9	79.8	87.7	71.4	81.2	71.4	81.6
BERT-P-EnA	85.3	91.2	80.5	87.9	71.7	81.3	71.3	81.4
BERT-EmA+BERT-P-EnA	85.8	91.4	80.7	88.0	72.3	81.5	71.7	81.5
BERT-EmA+BERT-P-EnA+CF (All)	**86.6***	**91.6***	**81.3***	**88.6***	**72.6***	**81.9***	**72.4***	**81.9***

Table 2: Performance of the proposed and previous methods on the Persona-Chat dataset under various persona configurations. The meanings of "Self Persona", "Partner Persona", "Original", and "Revised" can be found in Section 3. The results of P2 Bot was reported on the validation set. "-" denotes that the results were not reported in their papers. Numbers marked with * denote that the improvement over the best performing baseline is statistically significant (t-test with p-value < 0.05). Numbers in bold denote the combined fusion strategy that achieves the best performance.

pretrained using BERT on `Persona-Chat` dataset.

6.4 Experimental Results

Table 2 the evaluation results of our proposed and previous methods on `Persona-Chat` under various persona configurations. Our BERT-based model implemented with all the fusion strategies and concept encoding achieves a new state-of-the-art performance. We can see that incorporating the emotion and entailment knowledge of the utterances coupled with generic distributional semantics and external knowledge learned from pretraining rendered improvements on both **hits@1** and **MRR** conditioned on various personas. Compared to FT-PC (Mazaré et al., 2018) our best model outperformed it by 20.4 % in terms of hits@1 conditioned emotion, entailment and concepts. Compared to TransferTransfo (Wolf et al., 2019) and P^2 Bot (Liu et al., 2020) which were also trained using pretrained transformer models, our combined model outperformed them, which shows the effectiveness of fusion strategies and the concept-encoder. Lastly, our combined model outperformed the `BERT-CRA` (Gu et al., 2021b) in all the tasks. We see a 2.3 % and 1.9 % improvement in original and revised self-persona, and 1.4 % and 0.6 % improvement in original and revised partner-persona in terms of **hits@1**. The results bolster our hypothesis that emotion, entailment, and concepts play an important role in the task of response selection. Also, it is to be noted that `Persona-Chat` is a synthetic dataset, i.e. the data collection didn't happen naturally. Therefore, the chances are that the user will display this nuanced inter-play of persona and emotion is less. In addition to that, we observe the presence of contradictory distractor responses. Given this information, we see by introducing entailment aware fusion and concept encoding a significant performance improvement.

6.5 Human Evaluation

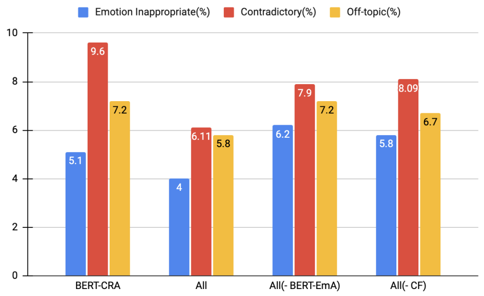


Figure 5: Human evaluation results on `Persona-Chat` self-persona original test split

Since a qualitative study by humans is necessary to understand the effectiveness of the proposed methods, we further perform human evaluation on a portion of the data. We randomly sampled 100 inferred examples from the test set by baseline model, combined model, combined model except the emotion, and combined model except the concept flow interaction. We combined all the samples and evaluated them using Amazon Mechanical Turk by 2 different turkers on three metrics: emotion inappropriate, contradictory and off-topic. The turkers needed to select if any of the three issues were present in an example. The percentages of reported

Models	hits@1	MRR
Baseline	84.4	90.7
BERT-EmA(− Speaker Encoding)	84.5	90.8
BERT-EmA	84.6	90.9
BERT-EnA-P	**85.3**	**91.2**

Table 3: Ablation Study for Emotion and Entailment on self original persona.

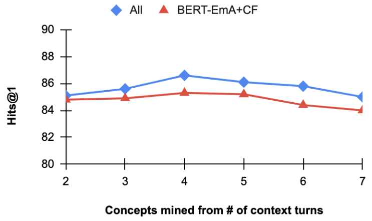


Figure 6: This graph shows how **hit@1** reaches an optimum value and then decreases with increase in number of turns used to mine concepts.

issues by both the groups is shown in Figure 5. The results reveal that all of our encoding pipelines are quite effective in reducing contradictory responses and somewhat effective in reducing off-topic and emotion inappropriate responses. The agreement observed between the two groups was in the moderate range (Krippendorff's $\alpha = 0.713$).

7 Analysis

7.1 Ablation Study for Emotion and Entailment

We perform ablation studies(shown in Table 3) to validate the effectiveness of emotion and entailment fusion in our proposed models. We see a very slight improvement in our baseline model that uses our proposed speaker embedding. Also, unsurprisingly effect of emotion is not that significant as the dataset is artificially created, but nonetheless some performance improvement is observed. Conditioning persona in entailment fusion improves the performance considerably as responses may not entail the persona of the speaker.

7.2 Effect of Context Turns on Concept Representation

Concept matching boosts the evaluation performance further. However, number of turns in the conversation history from which we mine the concepts influences the performances. It is evident from Figure 6 that most important concepts pertaining to the most relevant response will be present the recent conversation history.

personas	my favorite color is blue . I enjoy reading mysteries. I have seven children. I grew up on a large farm.
context	A: hello how are you today? B: I am well. how are you? A: I am doing great just got back from the beach B: that is great. I live far from the beach. A: I am very lucky we live beside the beach. what do you do for a living? B: I keep busy with my seven children. A: wow that much have taken some adjusting I teach kindergarten
golden response	do you reach mysteries to your children ? they are my favorite type of novel .
BERT-CRA	that must be a lot of work but very rewarding i bet
All	do you reach mysteries to your children ? they are my favorite type of novel .

Table 4: Case study showing concept flow.

7.3 Case Study

Table 4 shows the efficacy of concept-encoding, some times models fine-tuned on pretrained transformer models, like BERT-CRA tends to select a more generic responses rather than paying attention to the persona or specific keywords in the context. In this example, our proposed model better performs than BERT-CRA as it is conditioned on the concepts. Specifically, concepts in the correct response i.e "mysteries", "novel" relates to "reading mysteries" concept in the persona and "your children" relates to "teach kindergarten" in the context.

8 Conclusion

In this work, we propose a suite of novel fusion strategies and concept-flow encoder, which leverages emotion, entailment and concept information of the utterances. These features are not only helpful in improving the performances of our models but also provided key insights on certain aspects of how a humans communicate with each other. Though the techniques used in this paper is simple, it highlights the areas where response selection often falters, like detecting contraction, deviation from the concepts, etc. This work can be further extended by improving the concept representations using a graphical model.

References

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5935–5941, Online. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith

Ravi. 2020. Goemotions: A dataset of fine-grained emotions.

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

Jia-Chen Gu, Tianda Li, Zhen-Hua Ling, Quan Liu, Zhiming Su, Yu-Ping Ruan, and Xiaodan Zhu. 2021a. Deep contextualized utterance representations for response selection and dialogue analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2443–2455.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020a. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots.

Jia-Chen Gu, Tianda Li, Quan Liu, Xiaodan Zhu, Zhen-Hua Ling, Zhiming Su, and Si Wei. 2020b. Speaker-aware BERT for multi-turn response selection in retrieval-based chatbots. *CoRR*, abs/2004.03588.

Jia-Chen Gu, Zhen-Hua Ling, and Quan Liu. 2019a. Utterance-to-utterance interactive matching network for multi-turn response selection in retrieval-based chatbots.

Jia-Chen Gu, Zhen-Hua Ling, Xiaodan Zhu, and Quan Liu. 2019b. Dually interactive matching network for personalized response selection in retrieval-based chatbots. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1845–1854, Hong Kong, China. Association for Computational Linguistics.

Jia-Chen Gu, Zhenhua Ling, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2020c. Filtering before iteratively referring for knowledge-grounded response selection in retrieval-based chatbots. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1412–1422, Online. Association for Computational Linguistics.

Jia-Chen Gu, Hui Liu, Zhen-Hua Ling, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2021b. Partner matters! an empirical study on fusing personas for personalized response selection in retrieval-based chatbots. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 565–574, New York, NY, USA. Association for Computing Machinery.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring.

Mark R. Leary and Ashley Batts Allen. 2011. Personality and persona: Personality processes in self-presentation. *Journal of Personality*, 79(6):1191–1218.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Bill MacCartney and Christopher D. Manning. 2008. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 521–528, Manchester, UK. Coling 2008 Organizing Committee.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2020. I like fish, especially dolphins: Addressing contradictions in dialogue modeling.

Noah C. Pollock, Gillian A. McCabe, Ashton C. Southard, and Virgil Zeigler-Hill. 2016. Pathological personality traits and emotion regulation difficulties. *Personality and Individual Differences*, 95:168–177.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features.

Nadine R Richendoller and James B Weaver III. 1994. Exploring the links between personality and empathic response style. *Personality and individual Differences*, 17(3):303–311.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. *Automatic Keyword Extraction from Individual Documents*, chapter 1. John Wiley Sons, Ltd.

Yixuan Su, Deng Cai, Qingyu Zhou, Zibo Lin, Simon Baker, Yunbo Cao, Shuming Shi, Nigel Collier, and Yan Wang. 2021. Dialogue response selection with hierarchical curriculum learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1740–1751, Online. Association for Computational Linguistics.

Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric P. Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents.

Yuwei Wu, Xuezhe Ma, and Diyi Yang. 2021. Personalized response generation via generative split memory network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1956–1970, Online. Association for Computational Linguistics.

Chen Zhang, Hao Wang, Feijun Jiang, and Hongzhi Yin. 2021a. Adapting to context-aware knowledge in natural conversation for multi-turn response selection. In *Proceedings of the Web Conference 2021*, WWW '21, page 1990–2001, New York, NY, USA. Association for Computing Machinery.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too?

Zhuo Zhang, Danyang Zheng, and Ping Gong. 2021b. Multi-turn response selection in retrieval based chatbots with hierarchical residual matching network. *Journal of Physics: Conference Series*, 1757(1):012023.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566, Online. Association for Computational Linguistics.