Towards Faithful Personalized Response Selection in Retrieval Based Dialog Systems

Anonymous NAACL 2022 submission

Abstract

Personalized response selection systems are generally grounded on *persona*. However, the angle of emotion influencing response selection is not explored. Also, *faithfulness* to the conversation context of these systems 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 capture the interaction between persona, emotion, and entailment information of the utterances. A concept-flow encoder is designed which capture the relevant concept knowledge both in context and responses. 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 1.9% on original personas and 1.7% 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

With the advent of different natural language generation and understanding models, the open-domain conversational system has achieved great success (Roller et al., 2020; Komeili et al., 2021; Xu et al., 2021) and has found its application in various kinds of scenarios, ranging from personal assistants to social-bots (Saha et al., 2021; Konrád et al., 2021). Though the neural response generators improve the quality of responses significantly, however in many cases generated responses are not consistent with the persona of either the chatbot or the user, lacks emotion appropriateness, contradict themselves, go off-topic, etc. To overcome some of these shortcomings, many conversational systems employ a set of neural generators coupled with a re-ranking module (Saha et al., 2021; Konrád et al., 2021; Gao et al., 2020). Given a context, the job of this ensemble is to generate responses with different flavors and to select a response that is most relevant for that particular context.

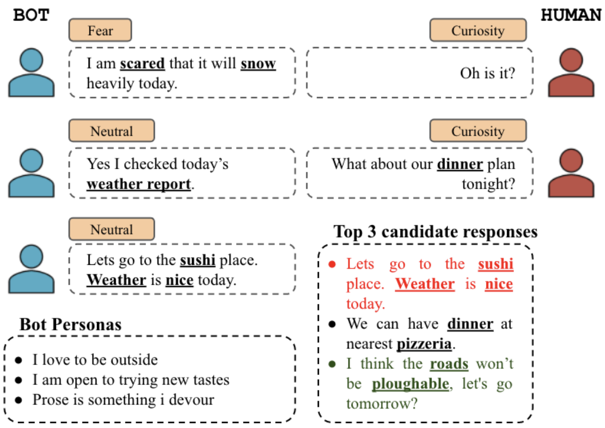


Figure 1: An example *unfaithful* response selection. For this conversation the selected candidate response directly contradicts the context. Also, the bot persona is influencing the response selection, while the situational emotions and concepts gets ignored. The underlines phrases/words denotes the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

Currently, most response selection systems are built in the context of information retrieval chatbots (Gu et al., 2021a; Zhang et al., 2021b; Gu et al., 2019a, 2020a). One of the issue with these systems is – they are trained on data where the diversity of responses are high. Secondly, these re-ranking systems have the 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 and reduces the *faithfulness* of the dialog system. Several works have achieved great success in incorporating persona while selecting(Gu et al., 2021b; Zhang et al., 2021a) or generating responses (Wu et al., 2021). However, no one has tried to incorporate emotion and persona interplay in response selection tasks. Figure 1 depicts situational emotion sometimes supersedes 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. We also observe concepts that are actively discussed in a conversation flow play an important role, and not much effort is made to incorporate this in response selection.

To increase the faithfulness and 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 speaker(bot) encoder to better represent the speaker-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 novel 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 1.9% 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 in Chatbots

Faithfulness in conversational systems 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 speaker 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 annotation.

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) release by AllenAI[1]. Second model is also a `RoBERTa` based model, a recently released NLI dataset, DECODE (Nie et al., 2020) is used for fine-tuning. 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. Also, we consider only `<contradiction>` and `<neutral>` labels. The usage of these labels varies depending on the model architecture.

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

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 hypothesis 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 $C_i = \{(U_j, E_j, ENTAIL_j)\}_{j=1}^{L}$ where U_j is an utterance, E_j is the set of emotions 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, $y_i \in \{0, 1\}$, $E_j \in \{0, 1, ...P\}$, and $ENTAIL_j \in \{$ `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).

5.2 Pretraining based models

The backbone framework used for different experiments is Bidirectional Encoder Representations from Transformers (BERT).

5.2.1 Speaker 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. Also, using BERT has its limitations, in some cases, the length of the input tokens often exceeds the maximum specified length for a model, which makes the overall context representation incomplete. To overcome this, (Gu et al., 2020b) introduced a speaker disentanglement strategy in form of *speaker embedding* fused with the original token representation. Though this technique has proven to improve response selection performance (Gu et al., 2020b; Su et al., 2021), however, the problem of maximum length truncation still exists. To circumvent this, we have created speaker-context encoding, which captures the representation of the speaker turns in the context while ignoring the listener's turns. The assumption here is, the speaker's turns will be most useful in selecting the relevant response. The input sequence that is sent to BERT to encode speaker context is composed as follows:

$$x_{si} = [CLS]u_{s1}[EOU]...u_{si}[EOU][SEP]r_i[EOU] \quad (1)$$

Where $u_{s1}, u_{s2}, ...u_{si}$ are speaker 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 $[CLS]$ token i.e. $\mathbf{h}^{\mathbf{s}}_{\mathbf{[CLS]}}$ are used in downstream tasks.

5.2.2 Baseline: Extension of `BERT-CRA`

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 instead of sequence A/B. Secondly, we fuse speaker-context encoding as described in the previous section with `BERT-CRA` encoding by doing multi-headed attention between the hidden representation of $[CLS]$ token of both encoder. The model is depicted in Figure 3b, mathematical representations are as follows:

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

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 last hidden states of $[CLS]$ token i.e. $\mathbf{h}^{\mathbf{CRA}}_{\mathbf{[CLS]}}$ 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 ($\mathbf{h}^{\mathbf{s}}_{\mathbf{[CLS]}}$, $\mathbf{h}^{\mathbf{CRA}}_{\mathbf{[CLS]}}$) and ($\mathbf{h}^{\mathbf{CRA}}_{\mathbf{[CLS]}}$, $\mathbf{h}^{\mathbf{s}}_{\mathbf{[CLS]}}$). For ease of presentation, we denote the whole multi-headed attention layer as $f_{mha}(*, *, *)$, $\mathbf{h_s}$ and $\mathbf{h_{CRA}}$ are the attention-layer outputs. Then these attention outputs are passed through an aggregation layer which basically concatenates $\mathbf{h_s}$ and $\mathbf{h_{CRA}}$ to get $\mathbf{h_i}$, finally the concatenated output is passed through a MLP to get the matching degree.

$$\begin{aligned} \mathbf{h_s} &= \mathbf{f_{mha}}(\mathbf{h}^{\mathbf{s}}_{\mathbf{[CLS]}}, \mathbf{h}^{\mathbf{CRA}}_{\mathbf{[CLS]}}, \mathbf{h}^{\mathbf{CRA}}_{\mathbf{[CLS]}}) \quad (3) \\ \mathbf{h_{CRA}} &= \mathbf{f_{mha}}(\mathbf{h}^{\mathbf{CRA}}_{\mathbf{[CLS]}}, \mathbf{h}^{\mathbf{s}}_{\mathbf{[CLS]}}, \mathbf{h}^{\mathbf{s}}_{\mathbf{[CLS]}}) \quad (4) \\ \mathbf{h} &= [\mathbf{h_s}; \mathbf{h_{CRA}}] \quad (5) \end{aligned}$$

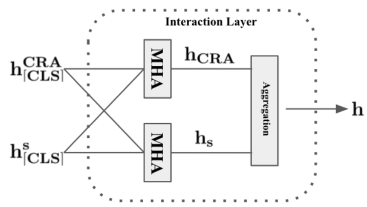


Figure 2: Interaction Layer

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:

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

5.2.3 `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 speaker-context as described in the previous section.

To incorporate emotion features in the BERT contextual representation, we attach emotion tags $[EMO_1]...[EMO_n]$ to each of the utterances, and each utterance can contain more than one emotion tag. The emotion-infused context representation is then concatenated with the original persona representation like as described in the previous section. 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:

$$\begin{aligned} x_{EmAi} = \; & [CLS]p_1p_2...p_i[EOP] \\ & [EMO_1]...[EMO_N]u_1[EOU] \\ & ...[EMO_1]...[EMO_n]u_i[EOU] \\ & [SEP]r_i[EOU] \end{aligned} \quad (7)$$

The rest of the architecture is the same as the baseline. The last hidden states of $[CLS]$ token is denoted by $\mathbf{h}^{\mathbf{EmA}}_{\mathbf{[CLS]}}$.

5.2.4 `BERT-EnA` : Entailment Aware Fusion

In this fusion strategy, the main objective is to model the entailment information about each of the speaker utterances with the response. Here, we are assuming that entailment information of listener utterances does not play a significant role in determining the correct response. Like `BERT-EmA` we follow a dual encoder pipeline, the first encodes the entailment information and the second encodes the speaker context. In this section, persona information is not taken into account.

To incorporate entailment features into BERT contextual representation, we attach entailment tags i.e. `<contradiction>` and `<neutral>` at the start of every speaker utterance. To maintain uniformity, we add a placeholder entailment tag `<neutral>` to the listener utterances. The response is concatenated with the utterance-entailment representation with a $[SEP]$ token. The input to entailment encoder is as follows:

$$\begin{aligned} x_{EmAi} = \; & [CLS][ENTAIL_i]u_{s1}[EOU] \\ & [ENTAIL_{neutral}]u_{l1}[EOU] \\ & ...[ENTAIL_{neutral}]u_{li}[EOU] \\ & [SEP]r_i[EOU] \end{aligned} \quad (8)$$

The rest of the architecture is the same as the baseline. The last hidden states of $[CLS]$ token is denoted by $\mathbf{h}^{\mathbf{EnA}}_{\mathbf{[CLS]}}$.

5.2.5 `BERT-EnA-P` : Persona-Entailment Aware Fusion

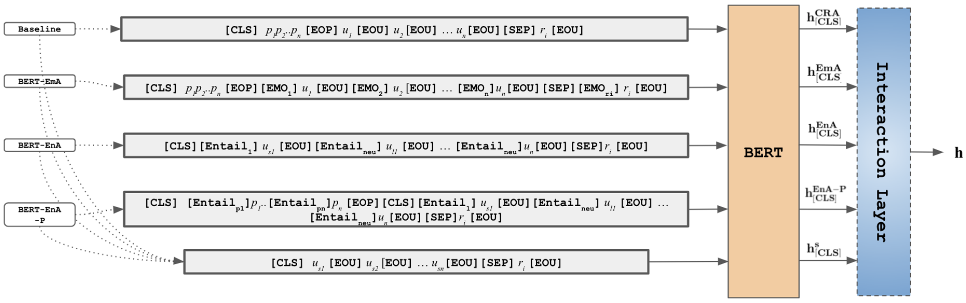
This is similar to `BERT-EnA`, the only difference is we are also attaching entailment information of each persona p_i with a given response r_i, the input to entailment encoder is as follows:

$$\begin{aligned} x_{EmA-Pi} = \; & [CLS][ENTAIL_i]p_1...[EOP] \\ & [ENTAIL_i]u_{s1}[EOU] \\ & [ENTAIL_{neutral}]u_{l1}[EOU] \\ & ...[ENTAIL_{neutral}]u_{li}[EOU] \\ & [SEP]r_i[EOU] \end{aligned} \quad (9)$$

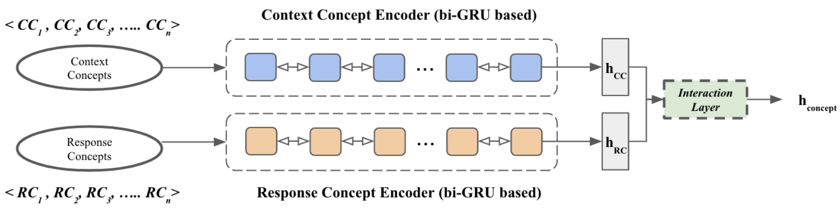
5.3 Concept-Flow(CF) Interaction

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 $\{\mathbf{ec_1}, \mathbf{ec_2}, ..., \mathbf{ec_n}\}$, , $\mathbf{ec_i} \in \mathbb{R}^{d_c}$ and $\{\mathbf{rc_1}, \mathbf{rc_2}, ..., \mathbf{rc_n}\}$, $\mathbf{rc_i} \in \mathbb{R}^{d_c}$



(a) A combined architecture of all the encoding pipelines, each of the pipeline combinations are shown by arrows. These encoding pipelines can be further combined with each other, combination strategies are explained in a later section.



(b) Concept-flow interaction network, the output of this network $h_{concept}$ can be concatenated with any of the BERT based encoder's output($\mathbf{h}$).

Figure 3: Overall Training Architecture. Though the BERT based encoders are independently depicted but it is trained along with concept-flow interaction(if included).

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 $\mathbf{h_i^{cc}}$, $\mathbf{h_i^{rc}}$ can be formulated as:

$$\mathbf{c_i^{cc}}, \mathbf{h_i^{cc}} = \overleftrightarrow{GRU}(\mathbf{ec_i}, \mathbf{h_{i-1}^{cc}}) \quad (10)$$

$$\mathbf{c_i^{rc}}, \mathbf{h_i^{rc}} = \overleftrightarrow{GRU}(\mathbf{er_i}, \mathbf{h_{i-1}^{rc}}) \quad (11)$$

$$\mathbf{h_{cc}} = tanh(\sum_{\mathbf{j \in 2*N_l}} \mathbf{W_j h_j^{cc}} + b) \quad (12)$$

$$\mathbf{h_{rc}} = tanh(\sum_{\mathbf{j \in 2*N_l}} \mathbf{W_j h_j^{rc}} + b) \quad (13)$$

Where $\mathbf{h_i^{cc}} \in \mathbb{R}^{2d_c}$, $\mathbf{h_i^{rc}} \in \mathbb{R}^{2d_c}$ are the i - the hidden states and $\mathbf{c_i^{cc}} \in \mathbb{R}^{2d_c}$, $\mathbf{c_i^{rc}} \in \mathbb{R}^{2d_c}$ are the outputs of the respective GRU encoders, $\mathbf{W_j}$ is a learn-able parameter and N_l is the number of layers in each GRUs. To model the interaction between $\mathbf{h_i^{cc}}$ and $\mathbf{h_i^{rc}}$ we follow the same interaction mechanism described in the earlier section. However, instead of concatenating the outputs from attention layers we sum them to reduce the computation time.

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`. 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 learn-

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	**86.2***	**91.6***	**81.1***	**88.5***	**72.5***	**81.8***	**72.3***	**81.9***

Table 1: 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.

ing 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 pretrained using BERT on `Persona-Chat` dataset.

6.4 Experimental Results

Table 1 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,

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	84.9	91
BERT-EnA-P	**85.3**	**91.2**

Table 2: Ablation Study for Emotion and Entailment on self original persona.

our combined model outperformed the `BERT-CRA` (Gu et al., 2021b) in all the tasks. We see a 1.9 % and 1.7 % improvement in original and revised self-persona, and 1.3 % and 0.5 % 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.

7 Analysis

7.1 Ablation Study for Emotion and Entailment

We perform ablation studies(shown in Table 2) 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 4 that most important concepts pertaining to the most relevant response will be present the recent conversation history.

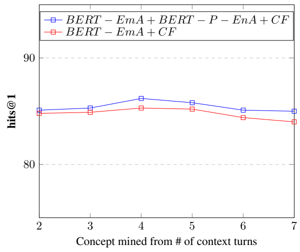


Figure 4: This graph shows how **hit@1** reaches an optimum value and then decreases with increase in number of turns used to mine concepts.

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
BERT-EmA +BERT-P-EnA+CF	do you reach mysteries to your children ? they are my favorite type of novel .

Table 3: Case study showing concept flow.

7.3 Case Study

Table 3 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.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data.

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.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2021. Internet-augmented dialogue generation.

Jakub Konrád, Jan Pichl, Petr Marek, Petr Lorenc, Van Duy Ta, Ondřej Kobza, Lenka Hýlová, and Jan Šedivý. 2021. Alquist 4.0: Towards social intelligence using generative models and dialogue personalization.

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.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot.

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

Sougata Saha, Souvik Das, Elizabeth Soper, Erin Pacquetet, and Rohini K. Srihari. 2021. Proto: A neural cocktail for generating appealing conversations.

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.

Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation.

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.