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-Chats (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-Chats 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-Chats 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-Chats 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-Chats 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-Chats (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-Chats 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 (McCartney and Manning, 2008) release by AllenAI.

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.

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 \( \sum_{j=1}^{N_{pk}} \frac{PMI(pk_j, rk_i)}{N_{pk}} \) and rejecting the concepts which are below a threshold value(\( \lambda \)). Similarly, for response/key 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 \( C_i = \{ \{U_j, E_j, ENTAIL_j\}\}_{j=1}^{M} \) where \( U_j \) is an utterance, \( E_j \) is the set of concepts present in \( U_j \) and \( ENTAIL_j \) is the entailment label of \( U_j \) with respect to \( r_i \) and \( y_i \). The \( j^{th} \) utterance \( U_j \) is denoted by \( U_j = \{u_{1j}, u_{2j}, ..., u_{Mj}\} \) 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:

$$x_{CRAi} = [CLS]p_1p_2...p_i[EOP]u_1[EOU]...u_i[EOU][SEP]r_1[EOU]$$ (2)

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. $h_{[CLS]}$ 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 $(h_{[CLS]}^s)$, $(h_{CRA}^s)$ and $(h_{[CLS]}^r)$, $(h_{CRA}^r)$ For ease of presentation, we denote the whole multi-headed attention layer as $f_{mha}(\ast, \ast, \ast)$, $h_s$ and $h_{CRA}$ are the attention-layer outputs. Then these attention outputs are passed through an aggregation layer which basically concatenates $h_{s}$ and $h_{CRA}$ to get $h_{1}$, finally the concatenated output is passed through a MLP to get the matching degree.

$$h_{s} = f_{mha}(h_{[CLS]}^{s}, h_{CRA}^{s}, h_{CRA}^{r}, h_{[CLS]}^{r})$$ (3)

$$h_{CRA} = f_{mha}(h_{[CLS]}^{CRA}, h_{s}^{CRA}, h_{[CLS]}^{CRA})$$ (4)

$$h = [h_{s}; h_{CRA}]$$ (5)

![Figure 2: Interaction Layer](image)

**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 \( \Theta \) be the model parameters then the loss function \( L(D, \Theta) \) for all the models can be formulated as follows:

\[
L(D, \Theta) = - \sum_{(C, p, r) \in D} y \log(g(C, p, r))
\]  

(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:

\[
x_{EmA_i} = [CLS]|ENTAIL_i|u_{i1}[EOU]
\]

\[
[ENTAIL_{neutral}]u_{i1}[EOU]
\]

\[
...[ENTAIL_{neutral}]u_{it}[EOU]
\]

\[
[SEP]r_{it}[EOU]
\]

(8)

The rest of the architecture is the same as the baseline. The last hidden states of \([CLS]\) token is denoted by \(h_{EmA}^{CLS}\).

### 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:

\[
x_{EnA-Pi} = [CLS]|ENTAIL_i|p_{1...}[EOU]
\]

\[
[ENTAIL_i|u_{i1}[EOU]
\]

\[
[ENTAIL_{neutral}]u_{i1}[EOU]
\]

\[
...[ENTAIL_{neutral}]u_{it}[EOU]
\]

\[
[SEP]r_{it}[EOU]
\]

(9)

The rest of the architecture is the same as the baseline. The last hidden states of \([CLS]\) token is denoted by \(h_{EnA}^{CLS}\).

### 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:

\[
x_{EnA-P_i} = [CLS]|ENTAIL_i|p_{1...}[EOU]
\]

\[
[ENTAIL_i|u_{i1}[EOU]
\]

\[
[ENTAIL_{neutral}]u_{i1}[EOU]
\]

\[
...[ENTAIL_{neutral}]u_{it}[EOU]
\]

\[
[SEP]r_{it}[EOU]
\]  

(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 \(\{ec_1, ec_2, ..., ec_n\}\), \(ec_1 \in \mathbb{R}^d\) and \(\{rc_1, rc_2, ..., rc_n\}\), \(rc_1 \in \mathbb{R}^d\).
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 $h^{ce}_i$, $h^{rc}_i$ can be formulated as:

$$c^{ce}_i, h^{ce}_i = GRU(e^{ce}_i, h^{ce}_{i-1})$$ (10)
$$c^{rc}_i, h^{rc}_i = GRU(e^{rc}_i, h^{rc}_{i-1})$$ (11)

$$h_{cc} = \tanh(\sum_{j \in 2:N_i} W_j h^{ce}_j + b)$$ (12)
$$h_{rc} = \tanh(\sum_{j \in 2:N_i} W_j h^{rc}_j + b)$$ (13)

Where $h^{ce}_i \in \mathbb{R}^{2d_e}$, $h^{rc}_i \in \mathbb{R}^{2d_e}$ are the $i$-th hidden states and $c^{ce}_i \in \mathbb{R}^{2d_c}$, $c^{rc}_i \in \mathbb{R}^{2d_c}$ are the outputs of the respective GRU encoders, $W_j$ is a learn-able parameter and $N_i$ is the number of layers in each GRUs. To model the interaction between $h^{ce}_i$ and $h^{rc}_i$ 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, $\lambda$ 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-
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 tasks.

  - **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 the Persona-Chat dataset under various persona configurations. The means 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Self Persona</th>
<th>Partner Persona</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original</td>
<td>Revised</td>
</tr>
<tr>
<td></td>
<td>hits@1</td>
<td>MRR</td>
</tr>
<tr>
<td>FT-PC (Mazaré et al., 2018)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DIM (Gu et al., 2019b)</td>
<td>78.8</td>
<td>86.7</td>
</tr>
<tr>
<td>TransferTransfo (Wolf et al., 2019)</td>
<td>80.7</td>
<td>-</td>
</tr>
<tr>
<td>P2 Bot (Liu et al., 2020)</td>
<td>81.9</td>
<td>-</td>
</tr>
<tr>
<td>FIRE (Gu et al., 2020c)</td>
<td>81.6</td>
<td>-</td>
</tr>
<tr>
<td>BERT-CRA (Gu et al., 2021b)</td>
<td>84.3</td>
<td>90.3</td>
</tr>
<tr>
<td>BERT-EmA</td>
<td>84.6</td>
<td>90.9</td>
</tr>
<tr>
<td>BERT-P-EnA</td>
<td>85.3</td>
<td>91.2</td>
</tr>
<tr>
<td>BERT-EmA+BERT-P-EnA</td>
<td>85.8</td>
<td>91.4</td>
</tr>
<tr>
<td>BERT-EmA+BERT-P-EnA+CF</td>
<td><strong>86.2</strong></td>
<td><strong>91.6</strong></td>
</tr>
</tbody>
</table>

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.
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.

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

<table>
<thead>
<tr>
<th>Models</th>
<th>hits@1</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>84.4</td>
<td>90.7</td>
</tr>
<tr>
<td>BERT-EnA−Speaker Encoding)</td>
<td>84.5</td>
<td>90.8</td>
</tr>
<tr>
<td>BERT-EnA</td>
<td>84.6</td>
<td>90.9</td>
</tr>
<tr>
<td>BERT-EnA-P</td>
<td>84.9</td>
<td>91</td>
</tr>
<tr>
<td><strong>BERT-EnA-P</strong></td>
<td><strong>85.3</strong></td>
<td><strong>91.2</strong></td>
</tr>
</tbody>
</table>

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 green eyes                                          | I have brown hair |
| I grew up in a large farm                                  | I grew up in a large city |

| context                                                                 |
|-------------------------------------------------------------------------|-------------------|
| A: hello how are you today?                                              | B: I am well. how are you? |
| B: I am doing great just got back from the beach                        | A: I am very lucky we live beside the beach |
| A: I am very lucky we live beside the beach                             | B: I keep busy with my seven children |
| A: wow that much has taken some adjusting                               | B: I keep busy with my seven children |

| golden response                                                          |
|-------------------------------------------------------------------------|-----------------|
| do you reach mysteries to your children ? they are my favorite type of novel | 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                           |
| BERT-EmA+BERT-P-EnA+CF                                                  | BERT-CRA |
| that must be a lot of work but very rewarding |

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


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


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?*