Guiding Topic Flows in the Generative Chatbot by Enhancing the ConceptNet with the Conversation Corpora

Anonymous ACL submission

Abstract

Human conversations consist of reasonable and natural topic flows, which are observed as the shifts of the mentioned concepts across utterances. Previous chatbots that incorporate the external commonsense knowledge graph prove that modeling the concept shifts can effectively alleviate the dull and uninformative response dilemma. However, there still exists a gap between the concept relations in the natural conversation and those in the external commonsense knowledge graph. Specifically, the concept relations in the external commonsense knowledge graph are not intuitively built from the conversational scenario but the world knowledge, which makes them insufficient for the chatbot construction. To bridge the above gap, we propose the method to supply more concept relations extracted from the conversational corpora and build an enhanced concept graph for the chatbot construction. We then introduce the enhanced graph to the response generation process with a designed network. Experimental results on the Reddit conversation dataset indicate our proposed method significantly outperforms strong baseline systems and achieves new SOTA results. Further analysis individually proves the effectiveness of the enhanced concept graph.

1 Introduction

With the rapid development of the natural language generation models (Radford et al., 2019; Zhang et al., 2020b; Brown et al., 2020) and the increase of the open-domain conversation corpora (Rashkin et al., 2019; Cui et al., 2020; Zhou et al., 2020; Zhang et al., 2018a), the quality of the response generated by the chatbot has been significantly improved. However, there still exist a series of challenges in the generative chatbot (Gao et al., 2019; Huang et al., 2020). Most of the time, users can still clearly distinguish between a human talker and a machine chatbot. Part of the reason is that the human is good at naturally switching the topics across the utterances, while the chatbot is relatively dull and tends to keep the topic still (Fang et al., 2018) or throw an unexpected topic (Wang et al., 2018; Tang et al., 2019).

As topic flows in the natural conversation could be observed as the shifts of the mentioned concepts across utterances, Zhang et al. (2020a) employ the ConceptNet (Speer et al., 2017) as the external knowledge graph and suggest that the graph provides relation-based one-hop and two-hop concepts to help the response generation. Their work is established on a restricted logical assumption: people would like to continuously talk on concepts that have commonsense relations to the current con-
cepts in the ConceptNet. We argue the assumption is too simple to imitate topic flows in human conversations. The ConceptNet is a commonsense graph built based on the concepts and their relations in the real world instead of in the natural conversational scenarios. Thus, only introducing the ConceptNet is insufficient for guiding the response generation. Figure 1 presents two instances in the Reddit conversation dataset for further explanation. Nodes and edges in the ConceptNet are marked to show concept shifts in conversations. For some concept relations that are common in the natural conversation, such as from “offline” to “internet” and from “Harden” to “rockets”, there are not corresponding edges in the ConceptNet. Therefore, only exploiting knowledge information in the ConceptNet could not cover topic flows in the natural conversation comprehensively.

To address the issue, we propose to construct an enhanced graph that consists of concept relations in both the commonsense knowledge graph and the natural conversation. Specifically, we extract new concepts as nodes and the high-frequency concurrence between concepts as edges from the conversation corpora. We then add these new nodes and new edges to the external knowledge graph to reconstruct the enhanced graph, which is used at the training and inference procedure for providing hints for the target response.

We then design a novel network to introduce the enhanced graph to the response generation. The experimental results on the Reddit conversation dataset show our method outperforms strong baselines and achieves new state-of-the-art performances on many metrics. We further conduct a series of analysis experiments, which results individually indicate the effectiveness of our proposed enhanced graph. Our contributions could be summarized in two folds, as follows:

- To bridge the gap between concept relations in the external knowledge graph and those in the natural conversation, we construct an enhanced graph with new nodes and edges extracted from the conversation corpora.
- Plenty of experiments verify the effectiveness of our method and the importance of concept relations in the conversation corpora. Our method achieves a new state-of-the-art performance on the Reddit conversation dataset.

## 2 Related Work

The end-to-end generative chatbot (Sutskever et al., 2014) achieves better performance in recent years due to more powerful model architectures (Radford et al., 2019; Zhang et al., 2020b; Brown et al., 2020) and larger conversation corpora (Zheng et al., 2019; Cui et al., 2020). However, there is still a series of challenges (Huang et al., 2020), such as off-topic and uninformative responses (Gao et al., 2019).

Aiming to give better responses, lots of works introduce external attributes into the response generation. Some of them introduce external attributes into the response generation. Zhou et al. (2018a) exploit concept relations in the ConceptNet to imitate concept shifts in human conversation. Zhang et al. (2020a) enhance the ConceptNet with the dialogue corpora to cover more human concept shifts. We propose to enhance the ConceptNet with the dialogue corpora to imitate topic flows better.

There also exist some works that focus on the dialogue relation extraction task. Some of them just get relationships among persons on a domain-specific dataset, instead of concept shifts on an open-domain dataset (Yu et al., 2020; Xue et al., 2021; Long et al., 2021). Others directly construct the conversation graph from the real conversation corpora for improving the response generation (Tang et al., 2019; Xu et al., 2020). However, their conversation graph only contains knowledge in the corpora, which quality is affected by the corpora.

## 3 Method

We present our method in this section. We first introduce the overview, then describe the pipeline of our method in detail.

### 3.1 Overview

Given a conversation corpus $D$ which contains many dialogue pairs such as $(X, Y)^1$, a human-
Figure 2: The pipeline of our method. It contains two parts. Firstly, we extract new nodes and new edges from the dialogue corpora, then merge them with the external graph to construct the enhanced graph. Secondly, we introduce the enhanced graph to the generation process. Specifically, we retrieval the subgraph in the enhanced graph according to the post $X$ and encode it with an improved Transformer architecture. Then, we apply the attention mechanism on the output of the Transformer architecture and the output of encoder to generate the response $Y$. The copy mechanism is also applied so that the response $Y$ could based on the subgraph directly.

3.2 Construct the Enhance Graph

We construct the enhanced graph $G_E$ on the basis of the external knowledge graph $G$ and the dialogue corpora, so that $G_E$ contains more knowledge of conversation topic flows than $G$. Formulating $G = \{V, E\}$ where $V$ and $E$ represent nodes and edges respectively, our method is extracting new nodes $V'$ and new edges $E'$ from the corpora, then reconstruct them into $G_E$. In other words, $G_E = \{V \cup V', E \cup E'\}$.

In order for new nodes to cover the conversation concepts as much as possible, we have two principles when extracting new nodes: common and concrete. Firstly, we set a frequency threshold to get common concepts according to $V$. Specifically, arranging the frequencies of $V$ in the dialogue corpora as $f_1, f_2, \cdots, f_{|V|}$, we set $f_{m \times |V|}$ as the threshold, and words which frequency higher than it are regarded as candidate concepts. Secondly, we choose nouns as new nodes from candidate concepts because nouns have rich semantic information than other types of words.!

We utilize the GIZA++ tool to extract (Och and 2005) (Och and 2003) (Och and 2002).

\[ Y' = \arg\max P(Y|X) \]
Ney, 2003) new edges, which represent topic flows in the conversation corpora. The GIZA++ tool is designed to align words in machine translation field. Its main idea is using the EM algorithm to iteratively train the bilingual corpus, and obtain word alignment from sentence alignment. We choose the toolkit here because we think concept alignment from source sentences to target sentences in the conversation is similar with bilingual word alignment. In practice, we first clean the corpora by removing all words except \(V \cup V'\). Then we run the GIZA++ toolkit to get the alignment probabilities. Finally, we arrange the probabilities to select the top \(k\) relations as new edges. Figure 3 presents an example. For the source concept “nurse”, we arrange all target concepts according to the alignment probabilities. The relations from “nurse” to the top \(k\) concepts are regarded as new edges, such as from “nurse” to “hospitical”. And we give these edges a new category: “DialogFlowTo”. Compared to other knowledge extraction method, our method adapts well to the parallel corpora of dialogue, and keeps interpretable and simple at the same time.

3.3 Response Generation

After building the enhanced graph \(G_E\), the next action is introducing it to the response generation process. This part is split into two steps: firstly, since introducing the whole graph to the generation process is unpractical and unnecessary, we retrieve a subgraph \(g\) from \(G_E\) and encode \(g\) based on an improved Transformer architecture. Secondly, we apply the attention mechanism and the copy mechanism to give the response based on \(g\).

Figure 4 presents how we encode the subgraph \(g\). Firstly, in order to model the interaction between the post \(X\) and the graph \(g\), a special node \(X'\) is added to \(g\) and connected to all nodes by encoding the post \(X\). We then alter the attention mask matrix, so that the target node could only get information from its source nodes. Specifically, if there is no edge \((a, b)\) from node \(a\) to node \(b\) in \(g\), we will mask the attention from \(b\) to \(a\). Finally, we introduce the edge type information to the vanilla Transformer architecture, because there are various types of edges in \(g\). And the forward calculation process of our improved architecture could be formulated as follows:

\[
h_{p}^{(l+1)} = FFN(h_{p}^{(l)} + u_{p}^{(l)}) \quad (2)
\]

\[
u_{p}^{(l)} = \sum_{q \in S(p)} a_{p,q}^{(l)} V^l(h_{q}^{(l)}) \quad (3)
\]

\[
a_{p,q}^{(l)} = Q^{(l)}(h_{p}^{(l)})K^{(l)}(h_{q}^{(l)})T + R^{(l)}(e_{q,p}) \quad (4)
\]

Where \(h_{p}^{(l)}\) is the vector of node \(p\) in the \(l\) layer, and \(u_{p}^{(l)}\) is information from source nodes of \(p\) in the \(l\) layer. \(S(p)\) is source nodes set of \(p\), and \(a_{p,q}^{(l)}\) is the attention weight. \(Q^{(l)}, K^{(l)}, V^l, R^{(l)}\) are different FFN networks in the \(l\) layer, and \(e_{q,p}\) is the type of edge \((q, p)\) \(4\).

The decoder generates the response \(Y\) based on \(g\). When generating \(t\)-th response token, the decoder state \(s_t\) is updated as follows:

\[
s_t = f_{dec}(s_{t-1}, y_{t-1}, v_{t-1}, e_{t-1}, \hat{e}_{t-1}^{graph}) \quad (5)
\]

\(4\)For edges from a node to itself, we give them a new category: “SelfTo”. For edges from and to \(X'\), we give them two new categories: “FromText” and “ToText”. 

![Figure 3: An example of the extract edges from the conversation corpora.](image)

![Figure 4: How we encode the subgraph. We add the special node \(X'\) to the graph by encoding the post \(X\). Then we improve the vanilla Transformer architecture to encode the graph. Attention mask corresponds to edges in the graph structure. And we also utilize edge type information in our architecture.](image)
Where $y_{t-1}$ is the token generated in the last step. $c_{t-1}^\text{ext}$ and $c_{t-1}^\text{graph}$ are outputs of the attention mechanism from the post and the subgraph, respectively. $f_{\text{dec}}$ are the updating function of the decoder. Besides generating tokens in the vocabulary, we also apply the copy mechanism so that the decoder could directly copy nodes from the subgraph $g$ as output tokens. We design a binary scalar $\sigma$ as a gate to control the generation source: vocabulary or $g$. In this way, the generation probability is the sum of probability on these two sources. The process could be formulated as follows:

$$
\sigma_t = \text{FFN}(s_t) \quad (6)
$$

$$
p_t = (1 - \sigma_t)p_t^{\text{voc}} + \sigma_t p_t^{\text{copy}} \quad (7)
$$

$$
p_t^{\text{voc}} = \text{FFN}_{\text{voc}}(s_t) \quad (8)
$$

$$
p_t^{\text{copy}} = \text{FFN}_{\text{graph}}(a_t) \quad (9)
$$

Where $p_t, p_t^{\text{voc}}$ and $p_t^{\text{copy}}$ are total prob, prob from vocabulary and prob from the subgraph, respectively. $\text{FFN}_{\text{voc}}$ and $\text{FFN}_{\text{graph}}$ are two linear networks and $a_t$ is the attention weight on the output of the improved Transformer architecture. The total loss of the generation process contains three parts: the generation loss, the copy loss, and the gate loss, as follows:

$$
\mathcal{L} = \mathcal{L}_{\text{gen}} + \mathcal{L}_{\text{copy}} + \mathcal{L}_{\text{gate}} \quad (10)
$$

$$
\mathcal{L}_{\text{gen}} = - \sum_t (1 - \sigma_t) \log p_t^{\text{voc}} \quad (11)
$$

$$
\mathcal{L}_{\text{copy}} = - \sum_t \sigma_t \log p_t^{\text{graph}} \quad (12)
$$

$$
\mathcal{L}_{\text{gate}} = - \sum_t \log \sigma_t + \log (1 - \sigma_t) \quad (13)
$$

### 4 Experiment

#### 4.1 Dataset

We conduct our experiments on Reddit conversation dataset (Zhou et al., 2018b), a single turn open-domain dialogue dataset which utterances are collected from Reddit. The dataset is large, containing 3,384,160 training pairs and 10,000 testing pairs. We utilize the preprocessed ConceptNet as the external knowledge graph (Speer et al., 2017), which includes 21,471 nodes and 120,850 edges. And there are 44 types of edges in the graph.

#### 4.2 Baselines

We follow Zhang et al. (2020a) and use three groups of models as baselines. We list them here:

- **Standard seq2seq model** (Sutskever et al., 2014): The model is based on the classical encoder-decoder framework. The encoder and the decoder are RNN architectures.
- **Knowledge enhanced seq2seq models**: MemNet(Ghazvininejad et al., 2018), CopyNet(Zhu et al., 2017), CCM(Zhou et al., 2018b) and ConceptFlow(Zhang et al., 2020a). These models introduce knowledge information into the generation process.
- **Pretrain Models**: GPT-2 lang(Zhang et al., 2020a), GPT-2 conv(Zhang et al., 2020a), DialoGPT(Zhang et al., 2020b). These models have a large number of parameters and have been pretrained on large corpus. GPT-2 lang and GPT-2 conv are built based on GPT-2(Radford et al., 2019).

For seq2seq, MemNet, CopyNet, CCM, GPT-2 lang and GPT-2 conv, we directly use results in ConceptFlow paper (Zhang et al., 2020a). For ConceptFlow, we run their public codes\(^5\). For DialoGPT, we finetune it on the dataset \(^6\).

#### 4.3 Evaluation Metrics

We use following metrics for evaluation:

- **Perplexity** (Serban et al., 2016): Perplexity measures the fluency of the responses.
- **Bleu** (Chen and Cherry, 2014), Nist (Doddington, 2002), ROUGE(Lin, 2004): These metrics measure the overlap between the generated responses and the ground truth.
- **Meteor** (Lavie and Agarwal, 2007): Meteor measure the relevance between the generated responses and the ground truth.
- **Entropy** (Zhang et al., 2018b): Entropy measures the diversity of generated responses.

We implement the above metrics based on the code of Galley et al. (2018)\(^7\).

#### 4.4 Implementation Details

During the process of constructing the enhanced graph, we utilize train dataset as the dialogue corpora. $m$ and $k$ are set to 20%, respectively. Since Zhang et al. (2020a) has processed the Reddit conversation dataset with the ConceptNet, we rebuild the dataset based on their data, and details could be found in the Appendix A.2. Table 1 presents

---


\(^6\) https://huggingface.co/microsoft/DialoGPT-medium

\(^7\) https://github.com/DSTC-MSR-NLP/DSTC7-End-to-End-Conversation-Modeling
the coverage of the ConceptNet and our enhanced graph on the Reddit conversation dataset.

For our model, we use two-layer GRUs (Cho et al., 2014) as the encoder and the decoder. We set the layers of the Transformer architecture to 3. We choose Adam as the optimizer, and the batch size, learning rate, max gradients norm, dropout are set to 30, 1e-4, 5, 0.2, respectively. We use TransE embedding (Bordes et al., 2013) and Glove embedding (Pennington et al., 2014) to initialize the embedding of concepts and words, respectively. We train our model on 8 V100 GPUs, and it takes about 1.5 hours to train an epoch. Our codes are presented in the supplementary materials.

<table>
<thead>
<tr>
<th>model</th>
<th>Bleu-3</th>
<th>Bleu-4</th>
<th>Nist-3</th>
<th>Nist-4</th>
<th>Rouge-1</th>
<th>Rouge-2</th>
<th>Rouge-L</th>
<th>meteor</th>
<th>PPL</th>
<th>Ent-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>seq2seq</td>
<td>0.0226</td>
<td>0.0998</td>
<td>1.1056</td>
<td>1.1069</td>
<td>0.1441</td>
<td>0.0189</td>
<td>0.1146</td>
<td>0.0611</td>
<td>48.79</td>
<td>7.6650</td>
</tr>
<tr>
<td>MemNet</td>
<td>0.0246</td>
<td>0.0112</td>
<td>1.1900</td>
<td>1.1977</td>
<td>0.1523</td>
<td>0.0215</td>
<td>0.1213</td>
<td>0.0632</td>
<td>47.38</td>
<td>8.4180</td>
</tr>
<tr>
<td>CopyNet</td>
<td>0.0226</td>
<td>0.0106</td>
<td>1.0770</td>
<td>1.0788</td>
<td>0.1472</td>
<td>0.0211</td>
<td>0.1153</td>
<td>0.0610</td>
<td>43.28</td>
<td>8.4220</td>
</tr>
<tr>
<td>CCM</td>
<td>0.0192</td>
<td>0.0084</td>
<td>0.9082</td>
<td>0.9095</td>
<td>0.1538</td>
<td>0.0211</td>
<td>0.1245</td>
<td>0.0630</td>
<td>42.91</td>
<td>7.8470</td>
</tr>
<tr>
<td>ConceptFlow</td>
<td>0.0495</td>
<td>0.0239</td>
<td>1.8838</td>
<td>1.8976</td>
<td>0.2241</td>
<td>0.0457</td>
<td>0.2032</td>
<td>0.0956</td>
<td>29.44</td>
<td>10.2390</td>
</tr>
<tr>
<td>GPT-2(lang)</td>
<td>0.0162</td>
<td>0.0162</td>
<td>1.0840</td>
<td>1.0844</td>
<td>0.1321</td>
<td>0.0117</td>
<td>0.1046</td>
<td>0.0637</td>
<td>29.08*</td>
<td>11.6500</td>
</tr>
<tr>
<td>GPT-2(lang)</td>
<td>0.0262</td>
<td>0.0124</td>
<td>1.1745</td>
<td>1.1763</td>
<td>0.1514</td>
<td>0.0222</td>
<td>0.1212</td>
<td>0.0629</td>
<td>24.55*</td>
<td>8.5460</td>
</tr>
<tr>
<td>DialoGPT</td>
<td>0.0189</td>
<td>0.0095</td>
<td>0.9986</td>
<td>0.9993</td>
<td>0.0985</td>
<td>0.0117</td>
<td>0.0971</td>
<td>0.0546</td>
<td>18.65*</td>
<td>9.8163</td>
</tr>
<tr>
<td>Ours</td>
<td>0.0644</td>
<td>0.0331</td>
<td>2.2573</td>
<td>2.2661</td>
<td>0.2592</td>
<td>0.0601</td>
<td>0.2340</td>
<td>0.1091</td>
<td>25.98</td>
<td>10.8173</td>
</tr>
</tbody>
</table>

Table 2: Evaluation results by human annotators. We also present Fleiss’ Kappa in the table. Kappa values range from 0.4 to 0.6, indicating fair agreement.

5.2 Human Evaluation

To evaluate model performances more comprehensively, we follow (Zhang et al., 2020a) and hire four human annotators to judge the quality of generated responses. Specifically, we sample 100 cases for three methods: ConceptFlow, ours, and ground truth responses. Annotators are required to score the responses from 1 to 3 on two aspects: fluency and appropriateness. Fluency evaluates whether a response is fluent or contains any grammar errors, while appropriateness evaluates whether a response is relevant to its post.

Human evaluation result is shown in Table 3. Obviously, ground truth responses get the highest scores. For entropy, our method gets the second-highest score, just lower than GPT-2. It proves that our proposed method could generate diverse responses. GPT-lang gets the highest diversity score, but it gets the lowest scores in most relevance metrics like Nist and Rouge. In comparison, our method has a good balance in relevance and diversity.
average scores. The average scores of our method are higher than the scores of ConceptFlow on both aspects, indicating our method could give more fluent and more relevant responses. And the best @1 ratios of our method are also higher than ConceptFlow, demonstrating that humans are more willing to chat with our chatbot.

The results of the automatic evaluation and human evaluation prove the effectiveness of our method. Based on the enhanced graph, our method could give responses of higher quality. Next, we conduct a series of experiments to study the effectiveness of the enhanced graph in detail.

### 5.3 Analysis

In this part, we conduct a series of experiments to study the effectiveness of the enhanced graph $G_E$.

**The enhanced graph VS the ConceptNet.** Considering that our method utilizes the enhanced graph and the improved Transformer architecture ($G_E + \text{Transformer}$) while ConceptFlow (Zhang et al., 2020a) utilizes the original ConceptNet and the GNN-based architecture named GRAFT-Net (Sun et al., 2018) ($G + \text{Transformer}$), we conduct two more models to directly compare $G_E$ and $G$.

The first model is built on $G +$ the improved Transformer, and the second is built on $G_E + \text{GRAFT-Net}$. The result is presented in Table 4. Obviously, with the same graph encoding architecture, models with $G_E$ achieve better performances on all metrics than models with $G$. The comparison results show that $G_E$ is more helpful to the response generation. And the importance of concept relations from the conversation corpora is also proved.

<table>
<thead>
<tr>
<th>model</th>
<th>Bleu-3</th>
<th>Bleu-4</th>
<th>Nist-3</th>
<th>Nist-4</th>
<th>Rouge-L</th>
<th>meteor</th>
<th>PPL</th>
<th>Ent-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours($G_E + \text{Transformer}$)</td>
<td>0.0644</td>
<td>0.0331</td>
<td>2.2573</td>
<td>2.2661</td>
<td>0.2340</td>
<td>0.1091</td>
<td>25.98</td>
<td>10.8173</td>
</tr>
<tr>
<td>$G + \text{Transformer}$</td>
<td>0.0615</td>
<td>0.0319</td>
<td>2.1448</td>
<td>2.1541</td>
<td>0.2307</td>
<td>0.1055</td>
<td>26.40</td>
<td>10.7081</td>
</tr>
<tr>
<td>$G_E + \text{GRAFT-Net}$</td>
<td>0.0529</td>
<td>0.0267</td>
<td>1.9270</td>
<td>1.9340</td>
<td>0.2115</td>
<td>0.0976</td>
<td>27.81</td>
<td>10.4316</td>
</tr>
<tr>
<td>ConceptFlow($G + \text{GRAFT-Net}$)</td>
<td>0.0493</td>
<td>0.0246</td>
<td>1.8265</td>
<td>1.8329</td>
<td>0.1888</td>
<td>0.0942</td>
<td>29.90</td>
<td>10.2700</td>
</tr>
</tbody>
</table>

Table 4: Evaluation results of models with different combinations of graphs and graph encoding architectures. The results show that $G_E$ outperforms $G$.

<table>
<thead>
<tr>
<th>model</th>
<th>Bleu-3</th>
<th>Bleu-4</th>
<th>Nist-3</th>
<th>Nist-4</th>
<th>Rouge-L</th>
<th>meteor</th>
<th>PPL</th>
<th>Ent-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>enhanced graph</td>
<td>0.0644</td>
<td>0.0331</td>
<td>2.2573</td>
<td>2.2661</td>
<td>0.2340</td>
<td>0.1091</td>
<td>25.98</td>
<td>10.8173</td>
</tr>
<tr>
<td>- edges in bottom 20%</td>
<td>0.0634</td>
<td>0.0328</td>
<td>2.2102</td>
<td>2.2194</td>
<td>0.2322</td>
<td>0.1070</td>
<td>27.17</td>
<td>10.7391</td>
</tr>
<tr>
<td>- edges in bottom 50%</td>
<td>0.0502</td>
<td>0.0249</td>
<td>1.8466</td>
<td>1.8528</td>
<td>0.2044</td>
<td>0.0938</td>
<td>30.77</td>
<td>10.2637</td>
</tr>
</tbody>
</table>

Table 5: Evaluation results after removing edges in the ConceptNet. More results are in the Appendix B.2

---

9We also compare the improved Transformer architecture and the GNN network. Because this is not the focus of this article, we write the results in the Appendix C.
Table 6: Three cases on the testset. We present responses generated by three different models. To study the impact of the knowledge graph, we mark concepts in the original ConceptNet in blue and concepts introduced by the enhanced graph in magenta.

5 and classify them into three categories roughly. The first type corresponds a pair of things that have a realistic relationship, such as “nurse” works for “hospital”. The second type corresponds a pair of things in the same kind, such as both “ps4” and “pc” are electronic devices. The third type corresponds a pairs of concepts with POS relationship, such as “perception” is the noun form of “perceptive”. These three categories are consistent with human common sense, proving our method could get various knowledge information from the real conversation corpora.

5.4 Case Study

To further study the improvement our method brings, we present three cases in Table 6. In case 1, DialoGPT and ConcpetFlow generate proper responses, but their responses are not as informative as ours. We could see that our response contains concept “episode” from $G_E$, and talks the same thing with the post. In case 2, it seems that DialoGPT and ConceptFlow don’t understand the post and give wrong responses. While our method gives high-quality response that contains concepts “source”, “server” and “bug”, which are relevant to the post. In case 3, for the post that about terrible football and super bowl, DialoGPT gives a short and dull response while ConceptFlow gives an unreasonable sentences. In contrast, our response is more consistent with the post. In summary, the enhanced graph $G_E$ could bring new concepts to the generated responses, and the responses generated based on $G_E$ are of higher quality. The result is consistent with automatic evaluation and manual evaluation.

Table 7: Concepts num in the generated responses.

<table>
<thead>
<tr>
<th>words num</th>
<th>concepts in $G_E$</th>
<th>concepts in $G$</th>
</tr>
</thead>
<tbody>
<tr>
<td>19.1056</td>
<td>2.2001</td>
<td>2.0593</td>
</tr>
</tbody>
</table>

Besides, we statistic the concepts in the generated responses on the testset, which is shown in Table 7. In generated response, there are 2.2 words in the enhanced graph $G_E$ on average. Compared to the ConceptNet, the enhanced graph indeed introduces new concepts into the responses. It also proves the effectiveness of our enhanced graph.

6 Conclusion

Because of the gap between the concept relations in the natural conversation and those in the external commonsense knowledge graph, just exploiting the knowledge information in the external knowledge graph is not sufficient to guide topic flows in the response generation. To address the issue, we propose to enhance the graph with knowledge in the dialogue corpora. We construct the enhanced graph and introduce it to the generation process with a designed network. Plenty of experiments on the Reddit dataset show our method outperforms other strong baselines and achieves new SOTA results. Further analysis indicates the effectiveness of the enhanced graph in detail. We will try to apply our proposed method to other domain-specific conversation datasets in the future.
References


Wenjie Wang, Minlie Huang, Xin-Shun Xu, Fumin Shen, and Liqiang Nie. 2018. Chat more: Deepening and widening the chatting topic via A deep model.

In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018, pages 255–264. ACM.


Wei-Nan Zhang, Qingfu Zhu, Yifa Wang, Yanyan Zhao, and Ting Liu. 2019a. Neural personalized response


A Data Processing

This part presents some details of data processing in this paper.

A.1 Extracting New Nodes and New Edges

As said in subsection 5.3, we conduct experiments to compare the effectiveness of topic flows in the external graph and those from the dialogue corpora. Besides extracting new edges, removing existing edges in the external graph is also based on the alignment probability. The process is shown in figure 6. If there exist edges from “nurse” to bottom n concepts in the ConceptNet, we will remove these edges. During our experiment, we set n to 20% and 50%, respectively.

A.2 Rebuild the Conversation Dataset

We conduct our experiments on Reddit conversation dataset (Zhou et al., 2018b). ConceptFlow (Zhang et al., 2020a) has processed the dataset with the ConceptNet. They get a subgraph for the post X, which contains 0-hop, 1-hop, and 2-hop nodes from source nodes N_x. Especially, they only keep 100 2-hop nodes in g and remove others.

For the fairness of the experiment, we rebuild the conversation dataset with the enhanced graph \( G_E \), based on their dataset. For the post X, we get a subgraph g in \( G_E \), and we present our method in Algorithm 1. Where \( V_0, V_1, V_2 \) are 0-hop, 1-hop, 2-hop nodes set, respectively. And \( V_{2-base} \) is the 2-hop nodes set in ConceptFlow dataset.

B Supplementary Evaluation Results

This part presents more evaluation results.

B.1 Supplementary Result for Overall Experiments

Table 8 shows supplementary evaluation result of generated responses. We use two new metrics for evaluation. Dist (Li et al., 2016) measures the diversity of generated responses, and ConceptPPL(Zhou et al., 2018b) calculates perplexity by considering both entities and words. We could see that our method gets the lowest Concept-PPL, showing the generated responses by our method are most fluent. Our method also achieves the best performances in Bleu and Nist, demonstrating that our method could give the most relevant responses. Pretrained models get the highest diversity scores because of the rich semantic information they get during the pretrain process. Besides these pretrained models, our method gets the highest diversity scores, showing our responses are the most informative. The supplementary result demonstrates that our method could give responses with higher quality than other baselines, and further confirms the effectiveness of the enhanced graph \( G_E \).
Table 8: Supplementary evaluation results on automatic metrics. We bold the best scores on each metric. Some models don’t utilize concept information, so Concept-PPL is not suitable for them.

<table>
<thead>
<tr>
<th>model</th>
<th>Bleu-1</th>
<th>Bleu-2</th>
<th>Nist-1</th>
<th>Nist-2</th>
<th>Dist-1</th>
<th>Dist-2</th>
<th>Concept-PPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>seq2seq</td>
<td>0.1702</td>
<td>0.0579</td>
<td>1.0230</td>
<td>1.0963</td>
<td>0.0123</td>
<td>0.0525</td>
<td>-</td>
</tr>
<tr>
<td>MemNet</td>
<td>0.1741</td>
<td>0.0604</td>
<td>1.0975</td>
<td>1.1847</td>
<td>0.0211</td>
<td>0.0931</td>
<td>46.85</td>
</tr>
<tr>
<td>CopyNet</td>
<td>0.1589</td>
<td>0.0549</td>
<td>0.9899</td>
<td>1.0664</td>
<td>0.0233</td>
<td>0.0988</td>
<td>40.27</td>
</tr>
<tr>
<td>CCM</td>
<td>0.1413</td>
<td>0.0484</td>
<td>0.8362</td>
<td>0.9000</td>
<td>0.0146</td>
<td>0.0643</td>
<td>39.18</td>
</tr>
<tr>
<td>ConceptFlow</td>
<td>0.2495</td>
<td>0.1064</td>
<td>1.6685</td>
<td>1.8531</td>
<td>0.0237</td>
<td>0.1268</td>
<td>26.76</td>
</tr>
<tr>
<td>GPT-2(lang)</td>
<td>0.1705</td>
<td>0.0486</td>
<td>1.0231</td>
<td>1.0974</td>
<td>0.0325</td>
<td>0.2461</td>
<td>-</td>
</tr>
<tr>
<td>GPT-2(conv)</td>
<td>0.1765</td>
<td>0.0625</td>
<td>1.0734</td>
<td>1.1623</td>
<td>0.0266</td>
<td>0.1218</td>
<td>-</td>
</tr>
<tr>
<td>DialoGPT</td>
<td>0.1404</td>
<td>0.0442</td>
<td>0.9195</td>
<td>0.9906</td>
<td>0.0632</td>
<td>0.2288</td>
<td>-</td>
</tr>
<tr>
<td>Ours</td>
<td>0.2872</td>
<td>0.1301</td>
<td>1.9607</td>
<td>2.2123</td>
<td>0.0256</td>
<td>0.1485</td>
<td>24.68</td>
</tr>
</tbody>
</table>

Table 9: Evaluation results of models when reducing edges in the ConceptNet.

<table>
<thead>
<tr>
<th>model</th>
<th>Bleu-1</th>
<th>Bleu-2</th>
<th>Nist-1</th>
<th>Nist-2</th>
<th>Rouge-1</th>
<th>Rouge-2</th>
<th>Dist-1</th>
<th>Dist-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>enhanced graph</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- edges in bottom 20%</td>
<td>0.2872</td>
<td>0.1301</td>
<td>1.9607</td>
<td>2.2123</td>
<td>0.2592</td>
<td>0.0601</td>
<td>0.0256</td>
<td>0.1485</td>
</tr>
<tr>
<td>- edges in bottom 50%</td>
<td>0.2821</td>
<td>0.1276</td>
<td>1.9234</td>
<td>2.1653</td>
<td>0.2591</td>
<td>0.0606</td>
<td>0.0251</td>
<td>0.1463</td>
</tr>
</tbody>
</table>

B.2 Supplementary Result for Experiments of Reducing Edges

We present the supplementary evaluation result of models when reducing edges in the ConceptNet in Table 9. Obviously, our method gets lower performances on almost all metrics, and removing 50% edges causes worse results than reducing 20% edges. Specifically, we find diversity scores drop a lot when reducing 50% edges. The result is consistent with the conclusion in subsection 5.3. The edges in the ConceptNet are also important and necessary for the response generation. And more concept flows helps to give more diverse and informative responses. The results above further prove that ConceptNet is vital for the generation process.

Table 10: Computation resources of different graph encoding architectures. Other modules of the network keep the same.

<table>
<thead>
<tr>
<th>model</th>
<th>parameters</th>
<th>training time/epoch</th>
</tr>
</thead>
<tbody>
<tr>
<td>improved Transformer</td>
<td>34.6M</td>
<td>1.5h</td>
</tr>
<tr>
<td>GRAFTGNN</td>
<td>35.3M</td>
<td>2.5h</td>
</tr>
</tbody>
</table>

C Analysis of the improved Transformer Architecture

In this part, we conduct a series of experiments to study the effectiveness of our proposed improved Transformer architecture.

C.1 The improved Transformer architecture VS the GRAFT-Net

From evaluation results in Table 4, we could see that with the same graph, models with the improved Transformer achieve higher scores on all metrics than models with the GRAFT-Net. The results demonstrate the improved Transformer could encode graphs better.

We also compare the parameters and training time of two architectures, which results are shown in Table 10. Obviously, our architecture contains fewer parameters with high training speed. The above two comparison shows the improved Transformer gets better performances than the GRAFT-Net while costing fewer computation resources.

C.2 Ablation study

We propose three improvements on vanilla Transformer architecture. To study the effectiveness of three improvements, respectively, we build corresponding ablation models, as follows:

- **w/o post node.** We remove the special node $X'$, and there is no interaction between the post $X$ and the subgraph $g$.
- **w/o edge mask.** We remove the edge mask, and the architecture is the vanilla Transformer.
- **w/o edge embed.** We remove the edge embedding in the architecture, and the edge type information is not introduced.

The evaluation results of these three ablation models are shown in Table 11. All ablation models get lower scores than the complete model on all metrics. The architecture without edge mask gets the lowest scores, indicating graph structure information in the knowledge graph is vital for the response generation and the vanilla Transformer architecture could not encode graph structures well. The results also prove the necessity of interaction between the
<table>
<thead>
<tr>
<th>model</th>
<th>Bleu-3</th>
<th>Bleu-4</th>
<th>Nist-3</th>
<th>Nist-4</th>
<th>Rouge-1</th>
<th>Rouge-2</th>
<th>Rouge-L</th>
<th>meteor</th>
<th>PPL</th>
<th>Ent-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>0.0644</td>
<td>0.0331</td>
<td>2.2573</td>
<td>2.2661</td>
<td>0.2592</td>
<td>0.0601</td>
<td>0.2340</td>
<td>0.1091</td>
<td>25.98</td>
<td>10.8173</td>
</tr>
<tr>
<td>w/o post node</td>
<td>0.0595</td>
<td>0.0305</td>
<td>2.1316</td>
<td>2.1402</td>
<td>0.2487</td>
<td>0.0562</td>
<td>0.2237</td>
<td>0.1044</td>
<td>27.00</td>
<td>10.7731</td>
</tr>
<tr>
<td>w/o edge mask</td>
<td>0.0573</td>
<td>0.0290</td>
<td>2.0694</td>
<td>2.0771</td>
<td>0.2442</td>
<td>0.0538</td>
<td>0.2201</td>
<td>0.1025</td>
<td>26.81</td>
<td>10.6822</td>
</tr>
<tr>
<td>w/o edge emb</td>
<td>0.0589</td>
<td>0.0295</td>
<td>2.1394</td>
<td>2.1472</td>
<td>0.2485</td>
<td>0.0547</td>
<td>0.2246</td>
<td>0.1050</td>
<td>26.46</td>
<td>10.6871</td>
</tr>
</tbody>
</table>

Table 11: Automation results of ablation models. All ablation models get lower scores than the complete model.

post and the subgraph, and the importance of the
edge type information.