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.

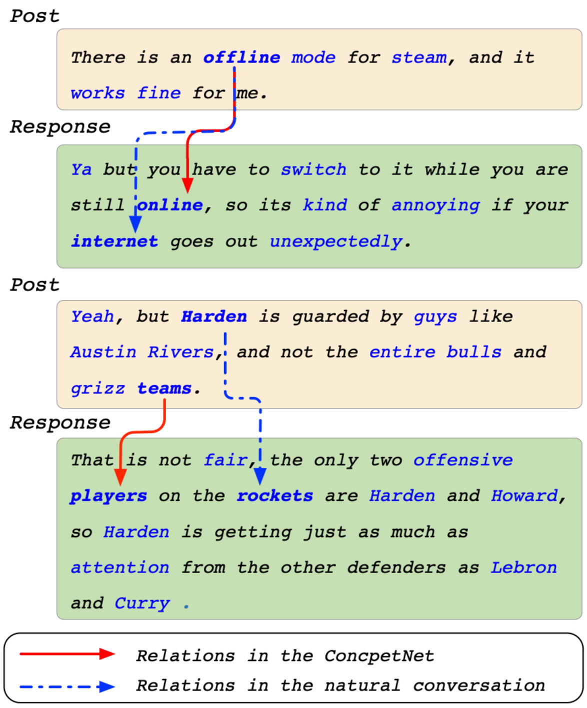

Figure 1: Two cases in the Reddit dataset. We use the ConceptNet as the external graph to show concept shifts in the conversation. Nodes are marked in blue. Concept relations in the graph and those in the natural conversation are marked with red solid lines and blue dashed lines, respectively.

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 concepts 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, like emotion(Zhou et al., 2018a; Wei et al., 2019), keywords (Xing et al., 2017; Wang et al., 2018) and persona (Zhang et al., 2018a, 2019a).

Besides, inspired by the fact that natural dialogue is based on human knowledge, plenty of previous work attempt to introduce external knowledge, such as the factual knowledge (Zhu et al., 2017; Ghazvininejad et al., 2018), the background document (Zhou et al., 2018c; Zhang et al., 2019b) and the commonsense knowledge graph (Zhang et al., 2020a; Zhou et al., 2018b) to the response generation. Zhou et al. (2018a) exploit concept relations in the ConceptNet to imitate concept shifts in human conversation. Zhang et al. (2020a) develops the idea further and utilizes the ConceptNet 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 a 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-

[1] We focus on single-turn conversations and leave multi-turn conversations to the future work.

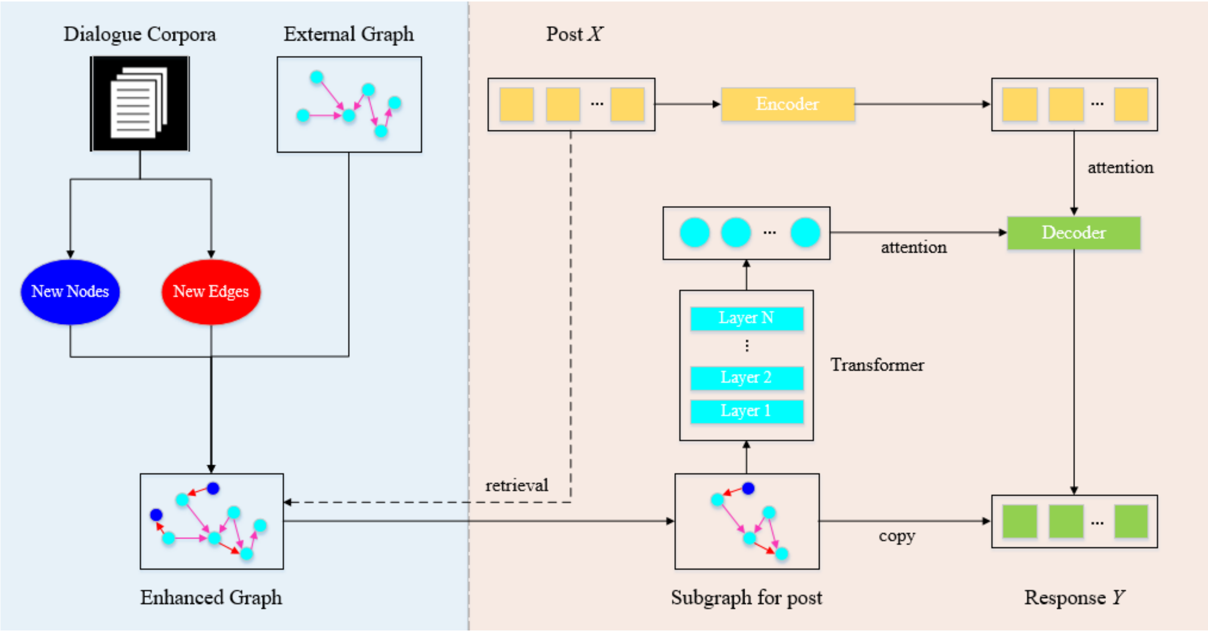

Figure 2: The pipeline of our method. It contains two parts. Firstly, we extract new nodes and new edges from the dialogue corpora, then merge them with the external graph to construct the enhanced graph. Secondly, we introduce the enhanced graph to the generation process. Specifically, we retrieval 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.

like chatbot is expected to generate the response Y based on the post X. The task could be formulated as generating best hypothesis Y' which maximizes the following probability:

$$Y' = argmaxP(Y|X) \tag{1}$$

Previous works introduce the external knowledge graph G to the task, aiming to give more coherent and informative responses. As expounded before, concept flows in common sense don't fit human conversations well. To address the issue, we enhance the external graph to help dialogue generation. Our method contains two parts, which is shown in Figure 2:

1. We enhance the external graph G with dialog corpora, and get an enhanced graph G_E. Specifically, We extract new edges and new nodes from the dialog corpora, then add them to G.
2. We introduce the enhanced graph G_E to the generation process, to improve the quality of responses. Firstly, we encode the graph with a designed Transformer structure. Then, the attention mechanism and the copy mechanism is applied to get information from the graph.

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 [2].

We utilize the GIZA++ tool to extract [3] (Och and

[2] We use the NLTK toolkit in python3 for POS tagging https://www.nltk.org/

[3] http://www.statmt.org/moses/giza/GIZA++.html

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.

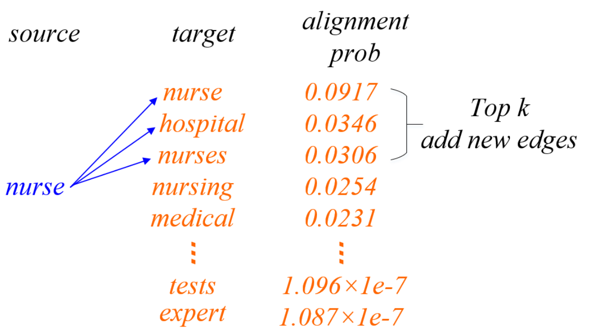

Figure 3: An example of the extract edges from the conversation corpora.

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:

$$h_p^{(l+1)} = FFN(h_p^{(l)} + u_p^{(l)}) \quad (2)$$

$$u_p^{(l)} = \sum_{q \in S(p)} a_{p,q}^{(l)} V^l(h_q^{(l)}) \quad (3)$$

$$a_{p,q}^{(l)} = Q^{(l)}(h_p^{(l)}) K^{(l)}(h_q^{(l)})^T + R^{(l)}(e_{q,p}) \quad (4)$$

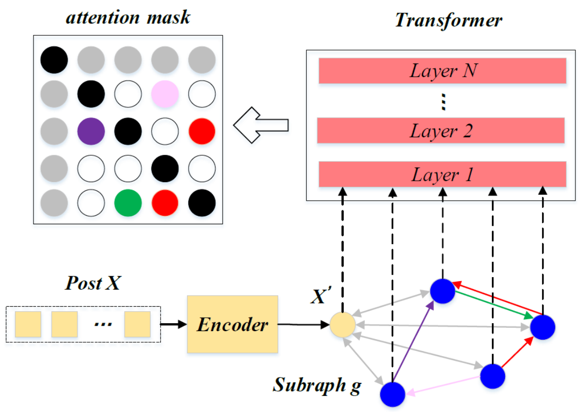

Figure 4: How we encode the subgraph. We add the special node X' to the graph by encoding the post X. Then we improve the vallina Transformer architecture to encode the graph. Attention mask corresponds to edges in the graph structure. And we also utilize edge type information in our architecture.

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:

$$s_t = f_{dec}(s_{t-1}, y_{t-1}, c_{t-1}^{text}, c_{t-1}^{graph}) \quad (5)$$

[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".

Where y_{t-1} is the token generated in the last step. c_{t-1}^{text} and c_{t-1}^{graph} are outputs of the attention mechanism from the post and the subgraph, respectively. f_{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 direct copy nodes from the subgraph g as output tokens. We design a binary scalar σ 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:

$$\sigma_t = FFN(s_t) \quad (6)$$

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

$$p_t^{vocab} = FFN_{vocab}(s_t) \quad (8)$$

$$p_t^{copy} = FFN_{graph}(a_t) \quad (9)$$

Where p_t, p_t^{vocab} and p_t^{copy} are total prob, prob from vocabulary and prob from the subgraph, respectively. FFN_{vocab} and FFN_{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:

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

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

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

$$\mathcal{L}_{gate} = -\sum_t \mathbb{I}_{y_t \in g} log \sigma_t + \mathbb{I}_{y_t \notin g} log(1-\sigma_t) \quad (13)$$

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.
- **Pretraind 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

[5] https://github.com/thunlp/ConceptFlow.
[6] https://huggingface.co/microsoft/DialoGPT-medium
[7] https://github.com/DSTC-MSR-NLP/DSTC7-End-to-End-Conversation-Modeling

graph	nodes	edges	response nodes	0-hop nodes		1-hop nodes		2-hop nodes	
				amount	golden	amount	golden	amount	golden
G	21471	120850	5.691	5.8129	0.5998	90.5138	1.2064	99.7706	0.8823
G_E	21754	218478	6.192	6.3223	0.6352	100.6227	1.4114	99.7706	0.8823

Table 1: Statistics of graphs coverage on the conversation dataset. Amount and golden are the number of total concepts and concepts appearing in responses, respectively. Obviously, G_E has a higher coverage than G.

model	Bleu-3	Bleu-4	Nist-3	Nist-4	Rouge-1	Rouge-2	Rouge-L	meteor	PPL	Ent-4
seq2seq	0.0226	0.0098	1.1056	1.1069	0.1441	0.0189	0.1146	0.0611	48.79	7.6650
MemNet	0.0246	0.0112	1.1960	1.1977	0.1523	0.0215	0.1213	0.0632	47.38	8.4180
CopyNet	0.0226	0.0106	1.0770	1.0788	0.1472	0.0211	0.1153	0.0610	43.28	8.4220
CCM	0.0192	0.0084	0.9082	0.9095	0.1538	0.0211	0.1245	0.0630	42.91	7.8470
ConceptFlow	0.0495	0.0239	1.8838	1.8896	0.2241	0.0457	0.2032	0.0956	29.44	10.2390
GPT-2(lang)	0.0162	0.0162	1.0840	1.0844	0.1321	0.0117	0.1046	0.0637	29.08*	**11.6500**
GPT-2(conv)	0.0262	0.0124	1.1745	1.1763	0.1514	0.0222	0.1212	0.0629	24.55*	8.5460
DialoGPT	0.0189	0.0095	0.9986	0.9993	0.0985	0.0117	0.0971	0.0546	**18.65***	9.8163
Ours	**0.0644**	**0.0331**	**2.2573**	**2.2661**	**0.2592**	**0.0601**	**0.2340**	**0.1091**	25.98	10.8173

Table 2: Evaluation results on automatic metrics. We bold the best scores on each metric. The PPL scores of pretrained models are not comparable because of different tokenization. The results indicate that our method gets the highest scores on most metrics. More results are in the Appendix B.1

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 method on 8 V100 GPUs, and it takes about 1.5 hours to train an epoch. Our codes are presented in the supplementary materials.

5 Evaluation

5.1 Automation Evaluation

The evaluation results are shown in Table 2. Except pretrain models, our method achieves the lowest PPL score, indicating that the responses generated by our model are more fluent. Bleu, Nist, Rouge, and meteor measure the relevance of generated responses and ground truth responses on different aspects. Our method outperforms all baselines by large margins on these metrics, demonstrating the responses generated by our method are more on-topic.

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.

5.2 Human Evaluation

	Fluency		
	Average	**Best @1**	**kappa**
ConceptFlow	2.2875	0.24	0.563
Ours	2.4325	0.30	0.603
Golden	**2.6975**	**0.69**	**0.665**
	Appropriateness		
	Average	**Best @1**	**kappa**
ConceptFlow	1.6200	0.12	0.480
Ours	1.6850	0.16	0.563
Golden	**2.3275**	**0.81**	**0.603**

Table 3: 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.

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 [8] . 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

[8](Zhang et al., 2020a) has proved that ConceptFlow outperforms a series of baselines. Therefore, we only use ConceptFlow as a comparison here in the case of limited human resources.

model	Bleu-3	Bleu-4	Nist-3	Nist-4	Rouge-L	meteor	PPL	Ent-4
Ours(G_E + Transformer)	0.0644	0.0331	2.2573	2.2661	0.2340	0.1091	25.98	10.8173
G + Transformer	0.0615	0.0319	2.1448	2.1541	0.2307	0.1055	26.40	10.7081
G_E + GRAFT-Net	0.0529	0.0267	1.9270	1.9340	0.2115	0.0976	27.81	10.4316
ConceptFlow(G + GRAFT-Net)	0.0493	0.0246	1.8265	1.8329	0.1888	0.0942	29.90	10.2700

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

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

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

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 + Transformer) while ConcpetFlow (Zhang et al., 2020a) utilizes the original ConceptNet and the GNN-based architecture named GRAFT-Net (Sun et al., 2018) (G + GRAFT-Net), 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 + 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 [9]. 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.

Concept relations extracted from the conversations corpora VS those in the ConceptNet. Now that we prove concept relations from the conversation corpora are important for the response generation, there is one more question to answer: Is it enough to only exploit concept relations from the conversation corpora? In other words, is the external commonsense knowledge graph such as the ConceptNet unnecessary? To study the question, we remove some edges in the ConceptNet when constructing the enhanced graph, and implementation details are given in the Appendix B.2. The evaluation result is shown in Table 5. Our method gets lower scores on all metrics after reducing edges. And reducing more edges results in worse performances. We could infer that concept relations, which are rare in the natural conversations, are also important for guiding topic flows in the response generation process. Therefore, knowledge information in both the external graph and the conversation corpora are necessary, and a good way is to merge them like our method.

related things		the same kind of things		form change	
nurse →	hospital	ps4 →	pc	perceptive →	perception
digest →	eat	jews →	arabs	piping →	pipe
rubble →	buildings	bling →	shiny	anarchist →	anarchy

Figure 5: Several examples of high-quality concept relations we extracted from the conversation corpora. We classify them into three categories.

Quality evaluation of the extracted edges. We conduct a human evaluation to verify the quality of the extracted edges. Specifically, we sample 100 extracted edges, and hire four human annotators to judge whether the target concept is relevant to the source concept. On average, 68 concept relations are marked as relevant edges. And there are 47 edges that all four annotators think relevant. We list some of these high-quality edges in Figure

[9] We 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.

#1 post	i drove home last night , saw my dad for the first time in 6th months , and slept all day today . woke up and poured myself a rum and coke thenstarted watching archer . today was a good day .
DialoGPT	i'm glad you had a good day.
ConcpetFlow	i 'm going to the same day . have a good night , man .
Ours	i 'm going to watch the first episode of archer and see if i can get a chance to sleep .
#2 post	what do you mean ? from the alpha or from the beta ? his uav was nerfed like crazy in the beta , not being able to ping or find a monster if they 're sneaking at all .
DialoGPT	i mean from the alpha. i'm not sure if it was nerfed in the beta, but i'm pretty sure it was.
ConcpetFlow	you can see the alpha of the alpha , but it 's a bug .
Ours	you can see the source on the server . i think he was just a bug .
#3 post	nowhere . fuck the patriots . fuck the seahawks . fuck football and fuck all these super bowl questions today .
DialoGPT	i'm with you man.
ConcpetFlow	you 're right . it 's a great way to get a new one . you can also see the seahawks and the seahawks .
Ours	how about you ask me to watch the game ? you want to watch the world cup and see how much of a rivalry it takes to get to the point ?

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.

words num	**concepts in G_E**	**concepts in G**
19.1056	2.2001	2.0593

Table 7: Concepts num in the generated responses.

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

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26-27, 2014, Baltimore, Maryland, USA*, pages 362–367. The Association for Computer Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734. ACL.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. Mutual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1406–1416. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.

Hao Fang, Hao Cheng, Maarten Sap, Elizabeth Clark, Ari Holtzman, Yejin Choi, Noah A. Smith, and Mari Ostendorf. 2018. Sounding board: A user-centric and content-driven social chatbot. In *Proceedings of NAACL-HLT 2018: Demonstrations*, pages 96–100, New Orleans, Louisiana.

Michel Galley, Chris Brockett, Xiang Gao, B. Dolan, and Jianfeng Gao. 2018. End-to-end conversation modeling : Moving beyond chitchat dstc 7 task 2 description (v 1 . 0).

Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019. Jointly optimizing diversity and relevance in neural response generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1229–1238. Association for Computational Linguistics.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5110–5117. AAAI Press.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Trans. Inf. Syst.*, 38(3):21:1–21:32.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119. The Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Xinwei Long, Shuzi Niu, and Yucheng Li. 2021. Position enhanced mention graph attention network for dialogue relation extraction. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 1985–1989. ACM.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

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

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5370–5381. Association for Computational Linguistics.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3776–3784. AAAI Press.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William W. Cohen. 2018. Open domain question answering using early fusion of knowledge bases and text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4231–4242. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric P. Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5624–5634. Association for Computational Linguistics.

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 Wei, Jiayi Liu, Xianling Mao, Guibing Guo, Feida Zhu, Pan Zhou, and Yuchong Hu. 2019. Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 1401–1410. ACM.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3351–3357. AAAI Press.

Jun Xu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Conversational graph grounded policy learning for open-domain conversation generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1835–1845. Association for Computational Linguistics.

Fuzhao Xue, Aixin Sun, Hao Zhang, and Eng Siong Chng. 2021. Gdpnet: Refining latent multi-view graph for relation extraction. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14194–14202. AAAI Press.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-based relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4927–4940. Association for Computational Linguistics.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020a. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2031–2043. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2204–2213. Association for Computational Linguistics.

Wei-Nan Zhang, Qingfu Zhu, Yifa Wang, Yanyan Zhao, and Ting Liu. 2019a. Neural personalized response generation as domain adaptation. *World Wide Web*, 22(4):1427–1446.

Yangjun Zhang, Pengjie Ren, and Maarten de Rijke. 2019b. Improving background based conversation with context-aware knowledge pre-selection. *CoRR*, abs/1906.06685.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018b. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1815–1825.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 270–278. Association for Computational Linguistics.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *CoRR*, abs/1901.09672.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018a. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 730–739. AAAI Press.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018b. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4623–4629. ijcai.org.

Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. Kdconv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7098–7108. Association for Computational Linguistics.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W. Black. 2018c. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 708–713. Association for Computational Linguistics.

Wenya Zhu, Kaixiang Mo, Yu Zhang, Zhangbin Zhu, Xuezheng Peng, and Qiang Yang. 2017. Flexible end-to-end dialogue system for knowledge grounded conversation. *CoRR*, abs/1709.04264.

A Data Processing

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

A.1 Extracting New Nodes and New Edges

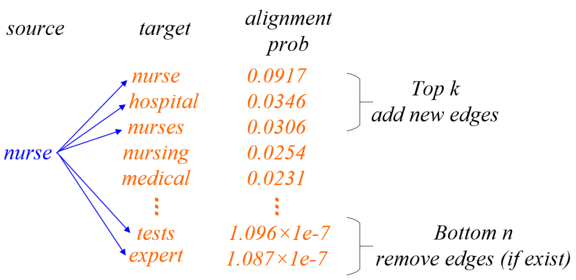

Figure 6: An example of the extract edges from the conversation corpus.

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.

Algorithm 1 Getting the subgraph g

Input: the post x and the enhanced graph G_E
Output: the subgraph g

```
Initiate V_g, E_g = ∅
Match x and V_e to get source nodes set V_x.
Initiate V_0 = V_x, V_1 = ∅, V_2 = ∅
for each node a ∈ V_0 do
    Get its neighborhood nodes set N_a ⊂ V_e.
    for each node b ∈ N_a do
        E_g = E_g ∪ {(a, b)}
        if b ∉ V_0 then
            V_1 = V_1 ∪ {b}
        end if
    end for
end for
for each node a ∈ V_1 do
    Get its neighborhood nodes set N_a ⊂ V_e.
    for each node b ∈ N_a do
        if b ∉ V_0 and b ∉ V_1 then
            if b ∈ V_{2-base} then
                V_2 = V_2 ∪ {b}
                E_g = E_g ∪ {(a, b)}
            end if
        else
            E_g = E_g ∪ {(a, b)}
        end if
    end for
end for
V_g = V_0 ∪ V_1 ∪ V_2
Return g = (V_g, E_g)
```

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 Concept-PPL(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 pretrianed 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.

model	Bleu-1	Bleu-2	Nist-1	Nist-2	Dist-1	Dist-2	Concept-PPL
seq2seq	0.1702	0.0579	1.0230	1.0963	0.0123	0.0525	-
MemNet	0.1741	0.0604	1.0975	1.1847	0.0211	0.0931	46.85
CopyNet	0.1589	0.0549	0.9899	1.0664	0.0233	0.0988	40.27
CCM	0.1413	0.0484	0.8362	0.9000	0.0146	0.0643	39.18
ConceptFlow	0.2495	0.1064	1.6685	1.8531	0.0237	0.1268	26.76
GPT-2(lang)	0.1705	0.0486	1.0231	1.0794	0.0325	**0.2461**	-
GPT-2(conv)	0.1765	0.0625	1.0734	1.1623	0.0266	0.1218	-
DialoGPT	0.1404	0.0442	0.9195	0.9906	**0.0632**	0.2288	-
Ours	**0.2872**	**0.1301**	**1.9607**	**2.2123**	0.0256	0.1485	**24.68**

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.

model	Bleu-1	Bleu-2	Nist-1	Nist-2	Rouge-1	Rouge-2	Dist-1	Dist-2
enhanced graph	0.2872	0.1301	1.9607	2.2123	0.2592	0.0601	0.0256	0.1485
- edges in bottom 20%	0.2821	0.1276	1.9234	2.1653	0.2591	0.0606	0.0251	0.1463
- edges in bottom 50%	0.2455	0.1055	1.6277	1.8144	0.2233	0.0476	0.0238	0.1262

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

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.

model	parameters	training time/epoch
improved Transformer	**34.6M**	**1.5h**
GRAFTGNN	35.3M	2.5h

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

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

model	Bleu-3	Bleu-4	Nist-3	Nist-4	Rouge-1	Rouge-2	Rouge-L	meteor	PPL	Ent-4
Ours	0.0644	0.0331	2.2573	2.2661	0.2592	0.0601	0.2340	0.1091	25.98	10.8173
w/o post node	0.0595	0.0305	2.1316	2.1402	0.2487	0.0562	0.2237	0.1044	27.00	10.7731
w/o edge mask	0.0573	0.0290	2.0694	2.0771	0.2442	0.0538	0.2201	0.1025	26.81	10.6822
w/o edge emb	0.0589	0.0295	2.1394	2.1472	0.2485	0.0547	0.2246	0.1050	26.46	10.6871

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.