# Enhancing Dialogue Generation with Conversational Concept Flows

**Siheng Li**[1]*, **Wangjie Jiang**[1]*, **Pengda Si**[1]*, **Cheng Yang**[1]
**Yao Qiu**[2], **Jinchao Zhang**[2], **Jie Zhou**[2], **Yujiu Yang**[1]†
[1]Shenzhen International Graduate School, Tsinghua University
[2]Tencent Inc, Beijing, China
{lisiheng21, jwj20, spd18}@mails.tsinghua.edu.cn
yang.yujiu@sz.tsinghua.edu.cn

## Abstract

Human conversations contain natural and reasonable topic shifts, reflected as the concept flows across utterances. Previous researches prove that explicitly modeling concept flows with a large commonsense knowledge graph effectively improves response quality. However, we argue that there exists a gap between the knowledge graph and the conversation. The knowledge graph has limited commonsense knowledge and ignores the characteristics of natural conversations. Thus, many concepts and relations in conversations are not included. To bridge this gap, we propose to enhance dialogue generation with conversational concept flows. Specifically, we extract abundant concepts and relations from natural conversations and build a new conversation-aware knowledge graph. In addition, we design a novel relation-aware graph encoder to capture the concept flows guided by the knowledge graph. Experimental results on the large-scale Reddit conversation dataset indicate that our method performs better than strong baselines, and further analysis verifies the effectiveness of each component.

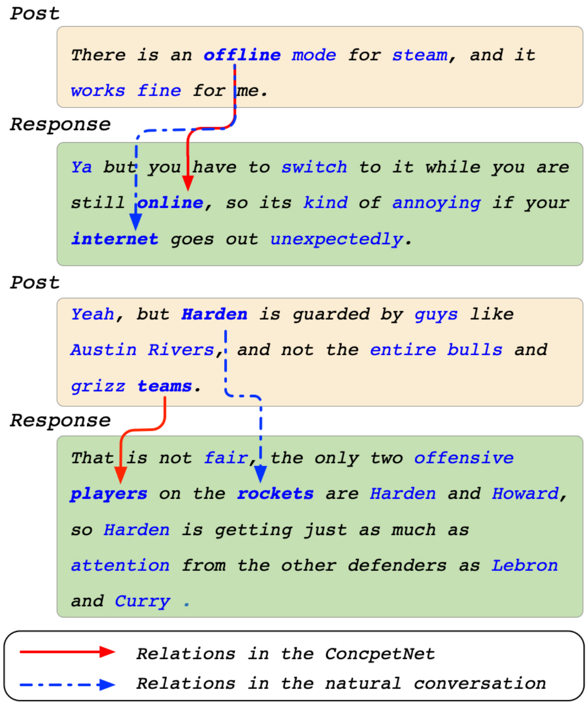


Figure 1: Two cases in the Reddit dataset. We use ConceptNet as the external knowledge graph to show concept flows in conversations. Concepts are marked in blue. Relations in the graph and those in the natural conversation are marked with red solid lines and blue dashed lines, respectively.

## 1 Introduction

With the remarkable development of conversation artificial intelligence (Shang et al., 2015; Adiwardana et al., 2020; Thoppilan et al., 2022), response generation has been improved in many ways, e.g., human-like persona (Zhang et al., 2018a), empathetic expression (Rashkin et al., 2019) and knowledge injection (Dinan et al., 2019), etc. However, there still exists a series of challenges (Gao et al., 2019; Xu et al., 2020a; Huang et al., 2020). One of the most noticeable is that humans are good at naturally switching topics during conversations, while machine-generated responses are relatively dull and tend to keep the topic still (Fang et al., 2018) or throw unexpected topics (Wang et al., 2018; Tang et al., 2019).

To overcome this challenge, previous works treat the topic shifts as concept flows (Zhang et al., 2020a; Zhou et al., 2018b, 2021a), which means traversing in the concept[1] space along relations in an external commonsense knowledge graph. Experimental results have shown that explicitly modelling concept flows effectively improves the relevance and engagingness of responses. However, we argue that there is a gap between the external knowledge graph and natural conversations. The most

* Equal contribution.
† Corresponding author (yang.yujiu@sz.tsinghua.edu.cn).

[1]Concept is the node in knowledge graph.



*Findings of the Association for Computational Linguistics: EACL 2023*, pages 1514–1525
May 2-6, 2023 ©2023 Association for Computational Linguistics

frequently used ConceptNet [2] (Speer et al., 2017) is limited to mostly (90%) taxonomic (e.g., *IsA*) or lexical (*e.g., Synonym*) knowledge, while contains relatively small portion of commonsense knowledge (Hwang et al., 2021). In addition, concepts and relations in natural conversations are more colloquial and timely. Thus, many concepts and relations are not included in the knowledge graph, which has also been verified in our experiments. As in Figure 1, the concept flows from "offline" to "internet" and from "Harden" to "rockets" are frequently observed in human conversations, while they are both not be included in the most frequently used ConceptNet.

To bridge the above gap and capture more concept flows, we propose to **E**nhance Dialogue Generation with **C**onversational **C**oncept **F**lows (**ECCF**). Specifically, we construct an enhanced knowledge graph that consists of concepts and relations in both commonsense knowledge graph and natural conversations. First, we extract new concepts as new nodes and the high-frequency relations between concepts as new edges from a large-scale dialogue corpora. Then, we add these new nodes and new edges to the commonsense knowledge graph to construct a **C**onverstaion-**A**ware **K**nowledge **G**raph (**CAKG**). To effectively guide concept flows in conversations with CAKG, we further propose a novel **R**elation-**A**ware **G**raph **E**ncoder (**RAGE**), which reasonably considers concepts and their relations in the graph encoding process for response generation.

We conduct a series of experiments on the large-scale Reddit conversation dataset (Zhou et al., 2018b; Baumgartner et al., 2020). Both automatic evaluation and human evaluation demonstrate that our method ECCF improves the relevance and diversity of responses, and outperforms strong baselines. Further analysis verifies the effectiveness of both CAKG and RAGE. Our research sheds light on explicitly modeling topic shifts with natural conversations.

## 2 Method

### 2.1 Overview

Given a dialogue context $X$, we aim to guide the topic shifts with the concepts and relations in a knowledge graph. Our method ECCF is shown in Figure 2, and can be summarized as follows:

1. Considering the abundant topic shifts in natural conversations, we enhance a commonsense knowledge graph $G$ with conversational concept flows extracted from large-scale conversation data. Then we get a conversation-aware knowledge graph $G_c$ (CAKG), which is more informative.

2. Fro response generation, we first encode the dialogue context $X$ with a context encoder. Then, to capture the concept flows defined in the knowledge graph $G_c$, we use a graph encoder for encoding the retrieved subgraph $g$ from $G_c$, which is based on the concepts in the dialogue context and their neighbor nodes. Last, we adopt a decoder with copy mechanism to generate a response and it can directly copy concepts from the subgraph $g$.

### 2.2 Knowledge Graph Enhancement with Conversational Concept Flows

We construct CAKG $G_c$ on the basis of the commonsense knowledge graph $G$ and a large-scale dialogue corpora Reddit (Baumgartner et al., 2020), so that $G_c$ contains more concept flows in natural conversation. Formulating $G = \{V, E\}$ where $V$ and $E$ represent nodes and edges respectively, we extract new nodes $V'$ and new edges $E'$ from the corpora, then reconstruct $G_c = \{V \cup V', E \cup E'\}$.

To obtain conversational concepts as much as possible, we have two principles when extracting new nodes: common and concrete. First, we set a frequency threshold $m$ and words with a frequency higher than it are regarded as candidate concepts. Second, we choose nouns as new nodes from candidate concepts because nouns have richer semantic information than other types of words [3].

We utilize the GIZA++ tool to extract [4] (Och and Ney, 2003) new edges, which represent concept flows in the conversations. The GIZA++ tool is designed to align words in the machine translation field. Its main idea is that utilize the EM algorithm to iteratively train the bilingual corpus and obtain word alignment from sentence alignment. We choose the toolkit here since concept alignments from source sentences to target sentences in

[2] ATOMIC (Sap et al., 2019) is also frequently used, while they focus more on human emotion and reaction in the generation of empathetic responses (Sabour et al., 2021; Tu et al., 2022), which we leave for future work.

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

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

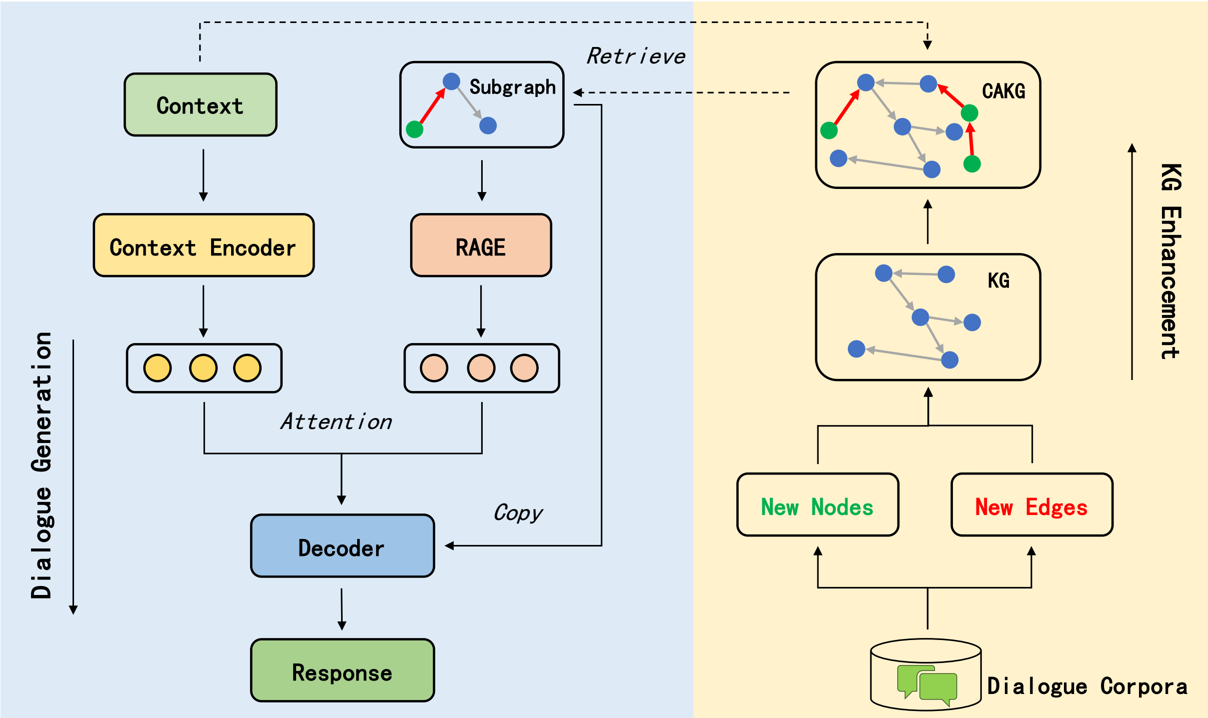


Figure 2: The pipeline of ECCF, which contains two parts. First, as in the right part, we extract new nodes and new edges from the dialogue corpora, then merge them with commonsense knowledge graph (KG) to construct conversational-aware knowledge graph (CAKG). Second, we use CAKG to guide the concept flows during the response generation process. For graph encoding, we use a relation-aware graph encoder (RAGE).

conversations are similar to 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 $k$ alignments as new edges. More details of the alignment process can be found in their original paper (Och and Ney, 2003).

An example is presented in Figure 3. For the source concept "nurse", we rank all the target concepts according to the alignment probabilities. The relations from "nurse" to the top $k$ concepts are regarded as new edges, such as "nurse $\rightarrow$ hospitical", and we attribute these edges to a new category: "DialogFlowTo".

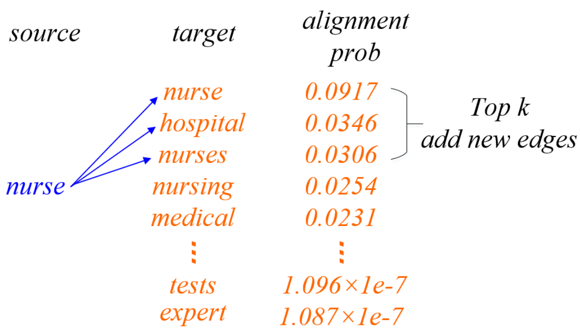


Figure 3: Extract concepts and relations from natural conversations.

### 2.3 Response Generation with Conversation-Aware Knowledge Graph

#### 2.3.1 Context Encoder

Given the dialogue context $X = (x_1, x_2, ..., x_m)$, we utilize a bi-directional encoder to get the contextual representation $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_m)$.

$$\mathbf{H} = \mathbf{Encoder}(X). \tag{1}$$

The encoder can be Transformer (Vaswani et al., 2017) or GRU (Cho et al., 2014), to be consistent with previous methods (Zhang et al., 2020a; Zhou et al., 2018b, 2021b), we utilize GRU in our experiments and choose the last word hidden states $\mathbf{h}_m$ as the representation of dialogue context.

#### 2.3.2 Relation-Aware Graph Encoder

Since introducing the whole graph to the generation process is unpractical and unnecessary, we retrieve a subgraph $g$ from $G_c$ and encode $g$ with the relation-aware graph encoder (RAGE), which is based on the Transformer Encoder (Vaswani et al., 2017). The subgraph $g$ derives from the concepts in the dialogue history and their one-hop and two-hop neighbor nodes[5]. To model the interactions between the dialogue context $X$ and subgraph $g$, we set a special node $\mathcal{X}$ to connect with all nodes of $g$, which represents the relations between dialogue and concepts. Then, we initialize the embedding of $\mathcal{X}$ with $\mathbf{h}_m$, and the embedding of $g$ with TransE embedding (Bordes et al., 2013). To model the graph structure of subgraph $g$, we design a graph mask matrix $M$:

$$m_{ij} = \begin{cases} 0 & \text{if } i = \mathcal{X} \text{ or } j = \mathcal{X}, \\ 0 & \text{if } i \in \text{Neighbor}(j), \\ -\infty & \text{otherwise}, \end{cases} \quad (2)$$

where $m_{ij} = 0$ indicates that node $i$ and node $j$ are connected, while $m_{ij} = -\infty$ represents the disconnect. Further, we replace the original Multi-Head Attention (MHA) with Relation-Aware Concept Attention (RACA), which incorporates the graph structure and node relations in the attention process. The differences are as follows:

$$\begin{aligned} \text{MHA} &= \text{softmax}(\frac{QK^T}{\sqrt{d}})V, \\ \text{RACA} &= \text{softmax}(\frac{QK^T}{\sqrt{d}} + M + R)V, \end{aligned} \quad (3)$$

where $Q$, $K$, $V$ is the query, key, and value vectors, more details in the original paper (Vaswani et al., 2017). $M$ represents the graph mask matrix and $R$ denotes edge relation bias:

$$r_{ij} = q^T \times e_{ij}, \quad (4)$$

where $e_{ij} \in \mathcal{R}^d$ is edge embedding [6], $q \in \mathcal{R}^d$ is used to transform the vector to scalar which represents relation importance in the attention process. We employ different $q$ in different heads and layers of the graph encoder, so that we can capture abundant and diverse relation-aware concept interactions. The output of the last layer is selected as the concept representations $\mathbf{G}$.

#### 2.3.3 Decoder

The decoder generates response $Y$ based on the dialogue context and subgraph. At $t$-th time step, the decoder state $s_t$ is updated as follows:

$$s_t = \textbf{Decoder}(s_{<t}, y_{t-1}, \mathbf{H}, \mathbf{G}) \quad (5)$$

[5] As the two-hop neighbor nodes are extensive, we select 100 two-hop nodes for each concept. For the fairness of the experiment, we use the same two-hop nodes set as in as in Zhang et al. (2020a).

[6] For the edges from a node to itself, we give them a new category: "SelfTO". For edges from and to $\mathcal{X}$, we give them two new categories: "FromText" and "ToText".

To be consistent with previous works, we utilize GRU in this paper. We employ attention mechanism to capture useful information from $\mathbf{H}$ and $\mathbf{G}$, more details in (Bahdanau et al., 2015).

In addition, we also apply the copy mechanism to directly copy concepts from subgraph $g$. The process can be formulated as follows:

$$\begin{aligned} \sigma_t &= \textbf{Sigmoid}(v_s^\top s_t), \\ p_t^v &= \textbf{Softmax}(\mathbf{W} \cdot s_t), \\ p_t^c &= \textbf{Softmax}(\mathbf{G} \cdot s_t), \\ p_t &= (1 - \sigma_t) \cdot p_t^v + \sigma_t \cdot p_t^c, \end{aligned} \quad (6)$$

where $p_t^v$ and $p_t^c$ are the probability of generation and copy, respectively.

#### 2.3.4 Objective Function

Our objective function has two parts, the first is the negative log likelihood of response generation:

$$\mathcal{L}_1 = -\sum_{t=1}^{n} \log p(x_t | x_{<t}, X, H, G). \quad (7)$$

We also supervise the copy gate as in Zhou et al. (2018a); Chen et al. (2022), so that the decoder can accurately copy concepts from the subgraph:

$$\mathcal{L}_2 = \sum_{t=1}^{n} q_t \cdot \log \sigma_t + (1 - q_t) \cdot \log(1 - \sigma_t), \quad (8)$$

where $q_t \in \{0, 1\}$ indicates whether $x_t$ is a concept word from the subgraph. The final objective function is $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$.

## 3 Experiment

### 3.1 Dataset

Follow Zhou et al. (2018b); Zhang et al. (2020a), we conduct experiments based on Reddit conversation dataset processed by (Zhou et al., 2018b). It contains 3,384,160 training pairs and 10,000 testing pairs. We use the commonsense knowledge graph ConceptNet (Speer et al., 2017) processed by Zhou et al. (2018b), which includes 21,471 nodes, 120,850 edges, and 44 types of edge relation.

### 3.2 Baselines

The baselines can be divided into three groups:

- **Standard seq2seq model**(Sutskever et al., 2014). The model is based on the classical encoder-decoder framework. The encoder and decoder are GRU as our model.

| 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 |
| ECCF | **0.0644** | **0.0331** | **2.2573** | **2.2661** | **0.2592** | **0.0601** | **0.2340** | **0.1091** | 25.98 | 10.8173 |

Table 1: Automatic Evaluations. We highlight the best scores on each metric. The PPL scores of pre-trained models are not comparable because of different tokenization. The results indicate that our ECCF gets the highest scores on most metrics.

- **Knowledge enhanced models**: MemNet(Ghazvininejad et al., 2018), CopyNet(Zhu et al., 2017), CCM(Zhou et al., 2018b) and ConceptFlow(Zhang et al., 2020a). These models explore knowledge information during 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[7]. For DialoGPT, we finetune it on the dataset [8].

### 3.3 Evaluation Metrics

We use the following metrics for evaluation:

- **PPL** (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 response and the ground truth.

- **Meteor** (Lavie and Agarwal, 2007): Meteor measures the relevance between generated responses and ground truth.

- **Entropy** (Zhang et al., 2018b): Entropy measures the diversity of generated responses.

[7]https://github.com/thunlp/ConceptFlow.

[8]https://huggingface.co/microsoft/DialoGPT-medium

We implement the above metrics based on the code of Galley et al. (2018) [9].

### 3.4 Implementation Details

For constructing CAKG, we utilize the training dataset for extracting conversational concept flows, which includes 3,384,160 utterance pairs. The frequency threshold $m$ is set as follows: we first arrange the frequencies of $V$ (original concepts in ConceptNet) in the dialogue corpora as $f_1, f_2, \cdots, f_{|V|}$, then, $f_{0.2\times|V|}$ is set as $m$. Noun words with frequency higher than $m$ is selected as new concepts. Further, we choose the top 20% concept relations for each concept as new edges.

For response generation, we use 2-layer GRU as context encoder and decoder, 3 layers of Transformer encoder with relation-aware concept attention as graph encoder. We choose Adam as the optimizer, the batch size, learning rate, max gradients norm, and 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 for one-epoch training.

## 4 Evaluation

### 4.1 Automation Evaluation

The experimental results are shown in Table 1. Except for pre-trained models, our method achieves the lowest PPL score, indicating that the responses generated by our model are more fluent. Furthermore, Bleu, Nist, Rouge, and Meteor measure the

[9]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 |
| $G_c$ | 21754 | 218478 | 6.192 | 6.3223 | 0.6352 | 100.6227 | 1.4114 | 99.7706 | 0.8823 |

Table 2: Statistics of graphs coverage on the conversation dataset. The amount and golden are the numbers of total concepts and concepts appearing in responses, respectively. Obviously, $G_c$ has higher coverage than $G$.

| | Fluency | | |
|---|---|---|---|
| | Average | Best @1 | kappa |
| **ConceptFlow** | 2.2875 | 0.24 | 0.563 |
| **ECCF** | 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 |
| **ECCF** | 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.

relevance between generated responses and ground truth responses in different ways. Our method outperforms all baselines by large margins on these metrics, demonstrating that the responses generated by our method are more relevant to the contexts and topic-consistent with humans. For diversity, our method gets the second-highest score, only lower than GPT-2. This proves that our proposed method can generate diverse responses. It is worth noticing that, although pre-trained models are slightly better at fluency and diversity, they perform much worse in relevance (Bleu, Nist, Rouge, Meteor) compared with our method and ConceptFlow. This indicates the superiority of explicitly modeling conversational topic shifts based on a knowledge graph.

### 4.2 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 randomly sample 100 cases for ConceptFlow, ours, and ground truth responses [10]. Annotators are required to score responses from 1 to 3 on two aspects: fluency and appropriateness. Fluency evaluates whether a response is fluent or contains grammar errors, while appropriateness measures whether a response is relevant and reasonable to its dialogue context.

As in Table 3, ECCF is better than the strong baseline ConceptFlow in terms of both fluency and appropriateness, the best @1 ratios of ECCF are also higher than ConceptFlow, demonstrating the superiority of our method. However, there is a large gap between ours and humans, indicating that there is still plenty of room for improvement.

## 5 Analysis

### 5.1 Conversation-Aware Knowledge Graph

Table 2 presents the statistics of ConceptNet $G$ and our CAKG $G_c$. Thanks to the conversational concept flows extracted from large-scale dialogue corpora, $G_c$ has more concepts and relations. Thus, more concepts in the responses are covered, especially for 0-hop and 1-hop concepts. This further proves the limitation of the external commonsense knowledge graph. We conduct an ablation study by replacing CAKG with ConceptNet (Ours w/o CAKG). As in Table 4, the performance drops in both relevance and diversity, which proves the effectiveness of conversational concept flows.

To further explore the relation between commonsense knowledge graph and conversational concept flows, we remove some edges in ConceptNet when constructing CAKG. As shown in Table 4, our method performs worse on relevance, fluency, and diversity, much worse when more edges are removed. Therefore, we can infer that concepts and relations in commonsense knowledge graph are also of great necessity for guiding topic flows in natural conversation. Further, both commonsense and conversation knowledge are beneficial to response generation, a reasonable way is to combine them as in our method.

### 5.2 Conversational Concept Flows

We conduct a human evaluation to verify the quality of the extracted conversational concept flows. Specifically, we randomly sample 100 extracted edges, and hire four human annotators to judge

[10] Zhang et al. (2020a) have proved that ConceptFlow outperforms a series of baselines including GPT-2 based methods. Therefore, we only use ConceptFlow for comparison here in the case of limited human resources.

| Model | Bleu-3 | Bleu-4 | Nist-3 | Nist-4 | Rouge-L | Meteor | PPL | Ent-4 |
|---|---|---|---|---|---|---|---|---|
| ECCF | 0.0644 | 0.0331 | 2.2573 | 2.2661 | 0.2340 | 0.1091 | 25.98 | 10.8173 |
| w/o CAKG | 0.0615 | 0.0319 | 2.1448 | 2.1541 | 0.2307 | 0.1055 | 26.40 | 10.7081 |
| w/o 20% edges in CN | 0.0634 | 0.0328 | 2.2102 | 2.2194 | 0.2322 | 0.1070 | 27.17 | 10.7391 |
| w/o 50% edges in CN | 0.0502 | 0.0249 | 1.8466 | 1.8528 | 0.2044 | 0.0938 | 30.77 | 10.2637 |
| w/o RAGE | 0.0529 | 0.0267 | 1.9270 | 1.9340 | 0.2115 | 0.0976 | 27.81 | 10.4316 |
| w/o graph mask | 0.0573 | 0.0290 | 2.0694 | 2.0771 | 0.2201 | 0.1025 | 26.81 | 10.6822 |
| w/o relation aware | 0.0589 | 0.0295 | 2.1394 | 2.1472 | 0.2246 | 0.1050 | 26.46 | 10.6871 |
| w/o dialogue node | 0.0595 | 0.0305 | 2.1316 | 2.1402 | 0.2237 | 0.1044 | 27.00 | 10.7731 |

Table 4: Analysis studies for conversation-aware knowledge graph (CAKG) and relation-aware graph encoder (RAGE), CN represents ConceptNet.

whether the target concept is relevant to the source concept. The results show that 68 edges are voted as relevant, of which 47 edges that all four annotators reach an agreement. According to our manually checking, these edges mainly have three categories, as shown in Figure 4. The first type corresponds to pairs that have realistic relations, such as "nurse" and "hospital". The second type corresponds to pairs in the same kind, such as both "ps4" and "pc" are electronic devices. The third type corresponds to pairs with POS relations, such as "perception" is the noun form of "perceptive". These three categories are meaningful, which proves that our method can obtain beneficial knowledge from natural conversations.

### 5.3 Relation-Aware Graph Encoder

We further investigate the effectiveness of the proposed relation-aware graph encoder (RAGE), and conduct several ablation studies as follows:

- **w/o RAGE.** To explore the superiority of our graph encoder, we replace it with a GNN-based architecture named GRAFT-Net (Sun et al., 2018), which is used by the strong baseline ConceptFlow (Zhang et al., 2020a).
- **w/o graph mask.** We remove the graph mask to explore the effectiveness of graph structure.
- **w/o relation aware.** We remove the relation bias in relation-aware concept attention, which aims to explore the effects of relation for graph encoding.
- **w/o dialogue node.** We remove the node $\mathcal{X}$ to study the necessity of the interactions between dialogue context and knowledge graph.

The results are shown in Table 4, and there are several findings. First, the performance drops largely when replacing our RAGE with traditional GNN used by previous work (Zhang et al., 2020a), which proves the effectiveness of transformer architecture and relation-aware concept attention. Second, the ablation study of each component further verifies the rationality of our design. Third, removing the graph mask results in worse performance compared with removing relation aware and removing dialogue node, which shows the importance of graph structure for graph encoding.

| 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 4: Several examples of high-quality concepts and relations we extracted from the conversation corpora. We classify them into three categories.

### 5.4 Case Study

To further investigate the performance, we present three cases in Table 5. In case 1, though the responses generated by DialoGPT and ConcpetFlow are proper, they are not as informative as ours. We can see that our response contains concept "episode" from $G_c$, which is consistent with the post and achieves a reasonable and natural topic shift. In case 2, it seems that DialoGPT and ConceptFlow don't understand the context and give wrong responses, while our method gives high-quality response that contains concepts "source", "server" and "bug", which are relevant to the context. In case 3, for the post about football and super bowl, DialoGPT gives a short and dull response while ConceptFlow gives an unreasonable sentence. In contrast, our response is more consistent with the post. These cases show that our method generates more suitable responses, which is consistent with the automatic evaluation and human evaluation.

| | | |
|---|---|---|
| **#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 . |
| ECCF | | 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 . |
| ECCF | | 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 . |
| ECCF | | 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 5: 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.

## 6 Related Work

**Dialogue Generation.** Recently, dialogue generation (Adiwardana et al., 2020; Thoppilan et al., 2022) has achieved great progress in many aspects. Pre-trained dialogue models (Zhang et al., 2020c; Roller et al., 2021) improve the response quality largely, even reaches human performance in single-turn dialogue generation. Persona-based dialogue system (Zhang et al., 2018a) possesses a human persona and is able to converse in a more captivating way. Rashkin et al. (2019) propose empathetic response generation, which aims to recognize partner feelings and reply accordingly. To bridge the gap between human utterances and dialogue system utterances, Chen et al. (2022) propose to enhance empathetic response generation with human-like intents. In this paper, we focus on the topic shifts during conversations and propose to enhance dialogue generation with conversational concept flows.

**Knowledge-Aware Dialogue Generation.** One of the most crucial challenges in dialogue generation is the lack of knowledge. Plentiful works have been proposed to inject reasonable knowledge into responses. One kind of these works utilizes unstructured knowledge, e.g., Wikipedia articles (Dinan et al., 2019), goal-related documents (Feng et al., 2021) etc. Another kind of work focuses on structured knowledge. Zhou et al. (2018a) exploit concept relations in commonsense knowledge graph to imitate concept shifts in human conversation. Zhang et al. (2020a) develop this idea and propose to explicitly model the concept flows in conversation. As we notice the gap between commonsense knowledge graph and natural conversations, we further propose to enhance dialogue generation with conversational concept flows.

There are also researches that extract information from natural conversations. Some of them extract relationships among persons on a domain-specific dataset (Yu et al., 2020; Xue et al., 2021; Long et al., 2021), while they focus on relation extraction not response generation. Others construct conversational graph from natural conversations to improve response generation (Xu et al., 2020b; Zou et al., 2021). However, their graphs only contain knowledge in conversations, while ignores the rich knowledge in commonsense knowledge graph. As shown in our analysis experiments, both types of knowledge are beneficial to response generation.

## 7 Conclusion and Future Work

In this paper, we argue the limitation of using external commonsense knowledge graph for response generation. To better capture topic shifts in natural conversation, we propose to enhance dialogue generation with conversational concept flows and construct conversation-aware knowledge graph. We further design a novel relation-aware graph encoder to capture the concept relations in knowledge graph. Extensive experiments on the large-scale Reddit dataset show the superiority of our method, and further analysis demonstrates the rationality of each component. In future work, we expect to capture more structural information from natural conversations to improve dialogue generation.

## Limitations

In this paper, we propose to enhance dialogue generation with conversational concept flows. Experimental results have shown that our method performs better than strong baselines. However, there are several major limitations. First, we use GIZA++ toolkit to extract concept relations, which is efficient but less expressive, as we cannot confirm the relations between concepts while they are quite different. For example, the relation between "nurse" and "hospital" is different to the relation between "thirsty" and "drink". These relations have certain semantics and can be beneficial for response generation. Second, the experimental results in this paper are only based on one dataset Reddit. Although Reddit is large and contains $3,384,160$ examples, more datasets can further verify the generalization ability of our methods. Third, we only combine conversational concept flows with ConceptNet (Speer et al., 2017), while other knowledge graphs (e.g., ATOMIC (Sap et al., 2019)) should be considered in future work to futher explore the relations between conversational concept flows and commonsense knowledge.

## Acknowledgements

This work was partly supported by the National Key Research and Development Program of China (No. 2020YFB1708200) and the Shenzhen Key Laboratory of Marine IntelliSense and Computation under Contract ZDSYS20200811142605016.

## References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pages 830–839. AAAI Press.

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.

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.

Mao Yan Chen, Siheng Li, and Yujiu Yang. 2022. Emphi: Generating empathetic responses with human-like intents. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1063–1074. Association for Computational 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.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. Multidoc2dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in*

*Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6162–6176. Association for Computational Linguistics.

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.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. 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 6384–6392. AAAI Press.

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*.

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.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 300–325. Association for Computational Linguistics.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2021. CEM: commonsense-aware empathetic response generation. *CoRR*, abs/2109.05739.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press.

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.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the*

*7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1577–1586. The Association for Computer Linguistics.

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.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. MISC: A mixed strategy-aware model integrating COMET for emotional support conversation. *CoRR*, abs/2203.13560.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

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.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020a. Recipes for safety in open-domain chatbots. *CoRR*, abs/2010.07079.

Jun Xu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020b. 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.

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.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020c. 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.

Hao Zhou, Minlie Huang, Yong Liu, Wei Chen, and Xiaoyan Zhu. 2021a. EARL: informative knowledge-grounded conversation generation with entity-agnostic representation learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2383–2395. Association for Computational Linguistics.

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.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2021b. Think before you speak: Using self-talk to generate implicit commonsense knowledge for response generation. *CoRR*, abs/2110.08501.

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.

Yicheng Zou, Zhihua Liu, Xingwu Hu, and Qi Zhang. 2021. Thinking clearly, talking fast: Concept-guided non-autoregressive generation for open-domain dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2215–2226. Association for Computational Linguistics.