CONVERSATION GENERATION WITH CONCEPT FLOW

Anonymous authors
Paper under double-blind review

ABSTRACT

Human conversations naturally evolve around related entities and connected concepts, while may also shift from topic to topic. This paper presents ConceptFlow, which leverages commonsense knowledge graphs to explicitly model such conversation flows for better conversation response generation. ConceptFlow grounds the conversation inputs to the latent concept space and represents the potential conversation flow as a concept flow along the commonsense relations. The concept is guided by a graph attention mechanism that models the possibility of the conversation evolving towards different concepts. The conversation response is then decoded using the encodings of both utterance texts and concept flows, integrating the learned conversation structure in the concept space. Our experiments on Reddit conversations demonstrate the advantage of ConceptFlow over previous commonsense aware dialog models and fine-tuned GPT-2 models, while using much fewer parameters but with explicit modeling of conversation structures.

1 INTRODUCTION

The rapid advancements of language modeling and natural language generation (NLG) techniques have enabled fully data-driven conversation models, which take user inputs (utterances) and directly generate natural language responses (Shang et al., 2015; Vinyals & Le, 2015; Li et al., 2016). On the other hand, the current generation models may still degenerate dull and repetitive contents (Holtzman et al., 2019; Welleck et al., 2019), which, in conversation assistants, lead to irrelevant, off-topic, and non-useful responses that would damage user experiences (Tang et al., 2019; Zhang et al., 2018; Gao et al., 2019).

A promising way to address this degeneration challenge is to model conversations with the help of knowledge, for example, open-domain knowledge graph (Ghazvininejad et al., 2018), commonsense knowledge base (Zhou et al., 2018a), or background documents (Zhou et al., 2018b). Recent research leverages these prior knowledge by grounding the conversation utterances to the external knowledge and integrating them as additional semantic representations; then response can be generated by conditioning on both the text inputs and the grounded semantics (Ghazvininejad et al., 2018; Zhou et al., 2018a;b).

Integrating external knowledge as a semantic representation of the utterance and an additional input to the conversation model effectively improves the quality of generated responses (Ghazvininejad et al., 2018; Logan et al., 2019; Zhou et al., 2018a). On the other hand, human conversations do not stay still on the same set of grounded semantics; instead, our dialog dynamically flows in the semantic space: we shift our discussions from one concept to another, chat about a group of related entities, and may switch dialog topics entirely (Fang et al., 2018). Limiting the usages of knowledge only to the grounded ones, effective as they are, does not leverage semantics' full potential in modeling human conversations.

This work presents ConceptFlow, (**Con**versation generation with Con**cept Flow**), which leverages the commonsense knowledge graph to model the conversation flow in the latent concept space. Given a conversation utterance, ConceptFlow starts from the grounded knowledge, which in our case are the commonsense concepts appearing in the utterance, and extends to multi-hop concepts along the commonsense relations. Then the conversation flow is modeled in the extended concept graph using a new fine-grained graph attention mechanism, which learns to encode the concepts using central or outer graph. Mimicking the development conversation topic flow, the graph attentions guild the concept flow by attending on different directions in the concept flow.

The encoded latent concept flow is integrated to the response generation with standard conditional language models: during decoding, each token, word or concept, is sampled from ConceptFlow's context vector, which combines the encodings of the utterance texts and the latent concept flow. This enables ConceptFlow to explicitly model the conversation structure when generating responses.

Our experiments on a Reddit conversation dataset (Zhou et al., 2018a) and a commonsense knowledge graph, ConceptNet (Speer & Havasi, 2012), demonstrate the advantage of ConceptFlow. In both automatic and human evaluation, ConceptFlow performs significantly better than various seq2seq based generation models (Sutskever et al., 2014), as well as previous methods that also leverage commonsense knowledge graph but as static memories (Zhou et al., 2018a; Ghazvininejad et al., 2018; Zhu et al., 2017). Notably, ConceptFlow also outperforms two fine-tuned GPT-2 systems (Radford et al., 2019), despite using much fewer parameters—Effective modeling of conversation structure can reduce the need of large parameter space.

We also provide extensive analyses and case studies to investigate the advantage of modeling conversation flow in the latent concept graph. Our analyses show that many of Reddit discussions are naturally aligned with the paths in the commonsense knowledge graph; expanding the latent concept graph multiple hops away from the initial grounded concepts significantly improves the coverage on the ground truth response. Our ablation study further confirms the effectiveness of our graph attention mechanism in selecting useful latent concepts and concepts appearing in golden responses, which help generate more relevant, informative, and less repetitive responses.

2 Methodology

This section presents our **Con**versation generation with latent Con**cept Flow** (ConceptFlow). As shown in Figure 1, the ConceptFlow models the conversation flow along commonsense relations between concepts to generate meaningful responses.

2.1 Preliminary on Grounded Conversation Models

Given a user utterance $X = \{x_1, ..., x_m\}$ with m words, conversation generation models often use an encoder-decoder architecture to generate a response $Y = \{y_1, ..., y_n\}$.

Typically, the **encoder** represents the user utterance X as a representation set $H = \{\vec{h}_1, ..., \vec{h}_m\}$. This is often done with a Gated Recurrent Unit (GRU):

$$\vec{h}_i = \text{GRU}(\vec{h}_{i-1}, \vec{x}_i), \tag{1}$$

where the $\vec{x}_i$ is the embedding of word x_i.

The **decoder** generates t-th response word according to the previous $t-1$ generated words $y_{<t} = \{y_1, ..., y_{t-1}\}$ and the user utterance X:

$$P(Y|X) = \prod_{t=1}^{n} P(y_t|y_{<t}, X). \tag{2}$$

The t-th token y_t is generated according to the t-step **decoder** context representation $\vec{s}_t$:

$$\vec{s}_t = \text{GRU}(\vec{s}_{t-1}, [\vec{c}_{t-1} \circ \vec{y}_{t-1}]), \tag{3}$$

where $\vec{c}_{t-1}$ is the context embedding of $t-1$-th time, $\vec{y}_{t-1}$ is the $t-1$-th generated word embedding and $\vec{s}_{t-1}$ is the decoder output representation of $t-1$-th time.

2.2 Conversation Generation with Latent Concept Flow

This part introduces the flow concept candidate construction, the latent concept flow encoding, and the conditional conversation decoder to generate response.

2.2.1 Constructing Flow Concept Candidates

ConceptFlow constructs a latent concept graph G for knowledge grounded conversation generation. The latent concept graph G starts from the grounded concepts (zero-hop concepts G^0), which appear in the conversation utterance and grounded by entity linking. Besides the grounded concepts,

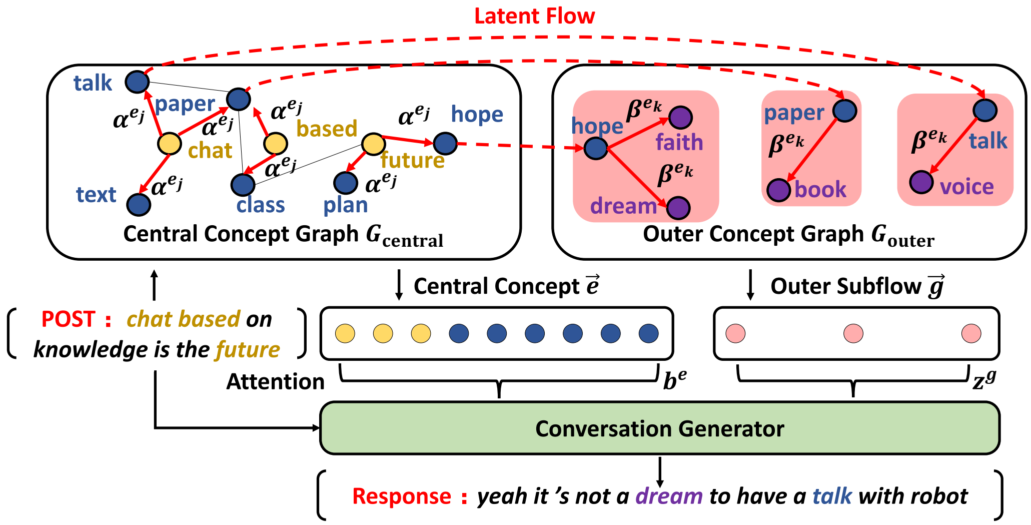

Latent Flow
talk
paper
α^{e_j}
chat
based
hope
future
text
class
plan
Central Concept Graph G_{central}
hope
β^{e_k}
faith
dream
paper
book
talk
voice
Outer Concept Graph G_{outer}
Central Concept $\vec{e}$
Outer Subflow $\vec{g}$
POST : chat based on knowledge is the future
Attention
b^e
z^g
Conversation Generator
Response : yeah it 's not a dream to have a talk with robot

Figure 1: The Architecture of ConceptFlow. The latent concept flow is annotated with red arrows.

ConceptFlow grows zero-hop concepts G^0 with one-hop concepts G^1 and two-hop concepts G^2. G^0 and G^1 form the central concept graph G_{central}, which is closely related to the current conversation topic. G^1 and G^2 construct an outer concept graph G_{outer}, which models outer conversation flow.

Latent concept flow consists of related concepts that can help understand the conversation. Next, we model conversation flow from zero-hop concepts, to one-hop concepts and then to two-hop concepts.

2.2.2 Encoding Latent Concept Flow

This part describes the latent concept flow encoding of central flow concepts and outer flow concepts.

Central Flow Encoding. The central flow concept models the concept flow from zero-hop concepts to one-hop concepts using the interactions between zero-hops and one-hops. A multi-layer Graph Neural Network (GNN) (Sun et al., 2018) is used to encode concept $e \in G_{\text{central}}$ in central concept graph. The l-th layer representation $\vec{e}_i^{\,l}$ of concept e_i is calculated by a single-layer feed-forward network (FFN) over three states:

$$\vec{e}_i^{\,l} = \text{FFN}\left(\vec{e}_i^{\,l-1} \circ \vec{p}^{\,l-1} \circ \sum_r \sum_{e_j} f_r^{e_j->e_i}\left(\vec{e}_j^{\,l-1}\right)\right), \tag{4}$$

where $\circ$ is the concatenate operator. $\vec{e}_j^{\,l-1}$ represents the concept e_j's representation of $(l-1)$-th layer. $\vec{p}^{\,l-1}$ represents user utterance representation of $(l-1)$-th layer.

The l-th layer user utterance representation is updated with the grounded concepts G^0:

$$\vec{p}^{\,l-1} = \text{FFN}\left(\sum_{e_i \in G^0} \vec{e}_i^{\,l-1}\right). \tag{5}$$

The $f_r^{e_j->e_i}(\vec{e}_j^{\,l-1})$ aggregates the concept semantics of the relation r specific neighbor concept e_j. It uses attention weight $\alpha_r^{e_j}$ to control concept flow from e_i:

$$f_r^{e_j->e_i}(\vec{e}_j^{\,l-1}) = \alpha_r^{e_j} \cdot \text{FFN}(\vec{r} \circ \vec{e}_j^{\,l-1}), \tag{6}$$

where $\circ$ is the concatenate operator and $\vec{r}$ is the relation embedding of r. The attention weight $\alpha_r^{e_j}$ is computed over all concept e_i's neighbor concepts according to the relation weight score and the Page Rank score (Sun et al., 2018):

$$\alpha_r^{e_j} = \text{softmax}(\vec{r} \cdot \vec{p}^{\,l-1}) \cdot \text{PageRank}(e_j^{\,l-1}), \tag{7}$$

where PageRank(e_j^{l-1}) is the page rank score to control propagation of embeddings along paths starting from e_i (Sun et al., 2018) and $\vec{p}^{l-1}$ is the $(l-1)$-th layer user utterance representation.

The 0-th layer concept representation $\vec{e}^0$ for concept e is initialized with the pre-trained concept embedding $\vec{e}$ and the 0-th layer user utterance representation $\vec{p}^0$ is initialized with the m-th hidden state h_m from the user utterance representation set H. The result GNN encodings establishes the central concept flow between zero-hop and one-hop concepts using attentions.

Outer Flow Encoding. The outer flow models the concept flow from one-hop concepts to two-hop concepts. Given a concept e_i from the one-hop concepts G^1, all neighbor concepts $e_k \in G^2$ are weighted to form the subflow g_{e_i}'s representation $\vec{g}_{e_i}$:

$$\vec{g}_{e_i} = \sum_{e_k} \beta^{e_k} \cdot [\vec{e}_i \circ \vec{e}_k], \tag{8}$$

where $\circ$ is the concatenate operator. The $\vec{e}_i$ and $\vec{e}_k$ are embeddings for e_i and e_k, respectively. The attention score β^{e_k} is calculated to weight and aggregate concept triple (e_i, r, e_k) to get $\vec{g}_{e_i}$:

$$\beta^{e_k} = \text{softmax}((w_r \cdot \vec{r})^\top \cdot \tanh(w_h \cdot \vec{e}_i + w_t \cdot \vec{e}_k)), \tag{9}$$

where r is the relation between the concept e_i and its neighbor concept e_k. The w_r, w_h and w_t are learnable parameters. The outer concept flow models more diverse developments of the conversation and guides the flow with the subgraph encoding to more possible directions.

2.2.3 Decoding from Latent Concept Flow

This part presents how to generate the response Y using the latent concept flow.

Context Representation with ConceptFlow. The t-th time output representation $\vec{s}_t$ of **decoder** is calculated by updating the $t-1$-th step output representation $\vec{s}_{t-1}$ with context representations:

$$\vec{s}_t = \text{GRU}(\vec{s}_{t-1}, [\vec{c}_{t-1} \circ \vec{y}_{t-1}]), \tag{10}$$

where $\vec{y}_{t-1}$ is the $t-1$-th step generated token y_{t-1}'s embedding. The context representation $\vec{c}_{t-1}$ is the concatenation of the text-based representation $\vec{c}_{t-1}^{\text{text}}$ and the concept-based context representation $\vec{c}_{t-1}^{\text{concept}}$:

$$\vec{c}_{t-1} = \text{FFN}([\vec{c}_{t-1}^{\text{text}} \circ \vec{c}_{t-1}^{\text{concept}}]). \tag{11}$$

$\vec{c}_{t-1}^{\text{text}}$ reads the user utterance representations H with the attention score a_{t-1}^j (Bahdanau et al., 2014):

$$\vec{c}_{t-1}^{\text{text}} = \sum_{i=1}^{m} a_{t-1}^j \cdot \vec{h}_j. \tag{12}$$

where the attention a_{t-1}^j weights over user utterance representations:

$$a_{t-1}^j = \text{softmax}(\vec{s}_{t-1} \cdot \vec{h}_j). \tag{13}$$

The concept-based representation $\vec{c}_{t-1}^{\text{concept}}$ is a combination of central concept flow encodings and outer flow encodings:

$$\vec{c}_{t-1}^{\text{concept}} = \sum_{e_i \in G_{\text{central}}} b_{t-1}^{e_i} \cdot \vec{e}_i \circ \sum_{g \in G_{\text{outer}}} z_{t-1}^{g} \cdot \vec{g}, \tag{14}$$

where the attention $b_{t-1}^{e_i}$ weights over central concept representations:

$$b_{t-1}^{e_i} = \text{softmax}(\vec{s}_{t-1} \cdot \vec{e}_i). \tag{15}$$

The attention z_{t-1}^g weights over outer flow representations:

$$z_{t-1}^g = \text{softmax}(\vec{s}_{t-1} \cdot \vec{g}). \tag{16}$$

Generating Words and Concepts. The conversation generator utilizes the t-th time output representation $\vec{s}_t$ to decode t-th words in response from the word vocabulary and the concept vocabulary:

$$y_t \sim \begin{cases} \text{softmax}(\vec{s}_t \cdot \vec{w}), \sigma^* = 0 \\ \text{softmax}(\vec{s}_t \cdot \vec{e}_i), \sigma^* = 1 \\ \text{softmax}(\vec{s}_t \cdot \vec{e}_k), \sigma^* = 2, \end{cases} \tag{17}$$

where $\vec{w}$ is the word embedding for word w, $\vec{e_i}$ is the central concept representation for concept e_i and $\vec{e_k}$ is the two-hop concept e_k's embedding.

The generation probability of word w is calculated over word vocabulary. The generation probability of concept is separated into two parts: central concept e_i's probability over $G^{0,1}$ and outer concept over G^2. The σ^* is a gate used to control the token generation from these three probability distributions:

$$\sigma^* = \text{argmax}_{\sigma \in \{0,1,2\}}(\text{FFN}_{\sigma}(\vec{s_t})), \tag{18}$$

to choose words ($\sigma^* = 0$), central concepts ($\sigma^* = 1$) and outer concepts ($\sigma^* = 2$) when generating the response.

Then we minimize the cross entropy loss L and optimize all parameters end-to-end:

$$L = \text{CrossEntropy}(y_t^*, y_t), \tag{19}$$

where the y_t^* is the ground truth tokens for words or concepts.

3 Experiment Settings

Dataset. All experiments use the Commonsense Conversation Dataset (Zhou et al., 2018a), which collects single-round dialogs from the Reddit site. This dataset contains 3,384,185 training pairs, 10,000 validation pairs and 20,000 test pairs. ConceptNet is used for our commonsense graph. It contains 120,850 triples, 21,471 concepts and 44 relations. For each example in the Commonsense Conversation Dataset, the average number of central concepts and two-hop concepts are 98.6 and 782.2, respectively.

Metrics. A wide range of evaluation metrics are included from three evaluating aspects: relevance, diversity and novelty. PPL (Serban et al., 2016), Bleu (Papineni et al., 2002), Nist (Doddington, 2002), ROUGE (Lin, 2004) and Meteor (Lavie & Agarwal, 2007) are used for relevance and repetitiveness; Dist-1, Dist-2 and Ent-4 are used for diversity, which is same with the previous work (Li et al., 2015; Zhang et al., 2018). Zhou et al. (2018a)'s concept PPL favors concept grounded models and is reported in Appendix A.1. The Precision, Recall and F1 Score to generate golden concepts (those appear in the ground truth response) are used to evaluate the quality of learned latent concept flow.

Baselines. Six baselines are compared in our experiments. Seq2Seq (Sutskever et al., 2014) is the basic encoder-decoder for the language generation task. MemNet (Ghazvininejad et al., 2018) and CopyNet (Zhu et al., 2017) utilize extra knowledge in two different ways: maintain a memory to store and read concepts; copy concepts for the response generation. Both MemNet and CopyNet provide solutions to store and incorporate knowledge for conversation generation. The Commonsense Knowledge Aware Conversation Generation Model (CCM) (Zhou et al., 2018a) leverages a graph attention mechanism to model local graphs, which further considers the graph structure for the improvement. The three models above use grounded graph as still knowledge and do not explicitly model conversation flow.

GPT-2 (Radford et al., 2019), the pre-trained model that achieves the state-of-the-art in lots of language generation tasks, is also compared in experiment. We fine-tune the 124M GPT-2 in two ways: concatenate all conversations together and train like a language model (GPT-2 (lang)); extend the GPT-2 model with encode-decoder architecture and supervised with response data (GPT-2 (conv)).

Implement Details. The zero-hop concepts are initialized by matching the keywords in post to concepts in ConceptNet, the same with CCM (Zhou et al., 2018a). Then zero-hop concepts are extended to their neighbors to form central concept graph. The outer concept flow usually contains lots of noise because of the large number of two-hop concepts. To select more related concepts for the conversation generation to reduce the computation cost, ConceptFlow first randomly selects 10% training data to train an initial version. Then we use the initial version's learned graph attention to select the top-100 two-hop concepts on all the rest data, and then conduct standard train, develop, and test process with the pruned graph. More details of this concept selection step can be found in Appendix C. TransE (Bordes et al., 2013) embedding and Glove (Pennington et al., 2014) embedding are used to initialize the representation of concepts and words, respectively. Adam optimizer with learning rate of 0.0001 is used to train the model.

Table 1: Relevance Between Generated and Golden Responses. The PPL results* of GPT-2 is not directly comparable because of the different vocabulary. More results can be found in Appendix A.1.

Model	Bleu-4	Nist-4	Rouge-1	Rouge-2	Rouge-L	Meteor	PPL
Seq2Seq	0.0098	1.1069	0.1441	0.0189	0.1146	0.0611	48.79
MemNet	0.0112	1.1977	0.1523	0.0215	0.1213	0.0632	47.38
CopyNet	0.0106	1.0788	0.1472	0.0211	0.1153	0.0610	43.28
CCM	0.0084	0.9095	0.1538	0.0211	0.1245	0.0630	42.91
GPT-2 (lang)	0.0162	1.0844	0.1321	0.0117	0.1046	0.0637	29.08*
GPT-2 (conv)	0.0124	1.1763	0.1514	0.0222	0.1212	0.0629	24.55*
ConceptFlow	**0.0246**	**1.8329**	**0.2280**	**0.0469**	**0.1888**	**0.0942**	**29.90**

Table 2: Diversity and Novelty of Generated Response. Diversity is calculated within generated responses; Novelty is compared to the input post. More results are in Appendix A.1.

	Diversity(↑)			Novelty w.r.t. Input(↓)				
Model	Dist-1	Dist-2	Ent-4	Bleu-4	Nist-4	Rouge-2	Rouge-L	Meteor
Seq2Seq	0.0123	0.0525	7.665	0.0129	1.3339	0.0262	**0.1328**	**0.0702**
MemNet	0.0211	0.0931	8.418	0.0408	2.0348	0.0621	0.1785	0.0914
CopyNet	0.0223	0.0988	8.422	0.0341	1.8088	0.0548	0.1653	0.0873
CCM	0.0146	0.0643	7.847	0.0218	**1.3127**	0.0424	0.1581	0.0813
GPT-2 (lang)	**0.0325**	**0.2461**	**11.65**	0.0292	1.7461	0.0359	0.1436	0.0877
GPT-2 (conv)	0.0266	0.1218	8.546	0.0789	2.5493	0.0938	0.2093	0.1080
ConceptFlow	0.0223	0.1228	10.27	**0.0126**	1.4749	**0.0258**	0.1386	0.0761

4 Evaluation

This section presents the quality of generated responses from the ConceptFlow, the ablation study for the roles of different modules, and case studies to evaluate the ConceptFlow.

4.1 Conversation Generation Quality Estimation

Automatic Evaluation. The relevance, diversity, and novelty of generated responses with different evaluation metrics are presented in Table 1 and Table 2.

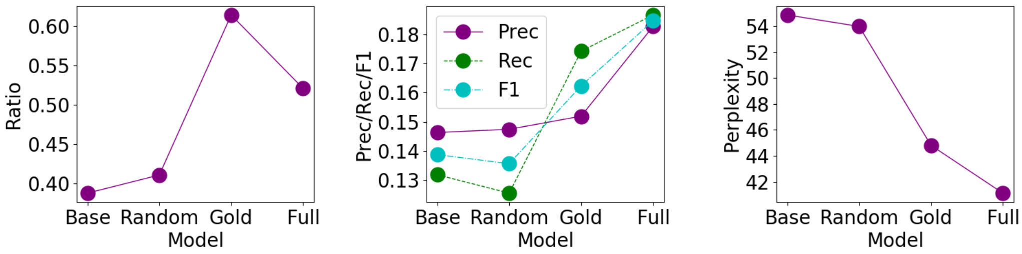
In Table 1, all the evaluation metrics compare the relevance between the generated response and the golden response. Our model outperforms all previous models by large margins. Responses generated by our model are on-topic and cover more necessary information. In Table 2, Dist-1, Dist-2, and Ent-4 measure the word diversity of generated response, whereas the rest of metrics measure the repetitiveness comparing to the user utterance to avoid dull copying the input. Our model also presents a convincing balance to generate novel and diverse responses. GPT-2 (lang) performs more diversely, but ConceptFlow performs more novelty and more on-topic than both GPT-2 versions, perhaps due to its different decoding strategy.

Human Evaluation. The human evaluation mainly focuses on two testing scenarios: appropriateness and informativeness, which are important for conversation systems. Appropriateness indicates if the response is on-topic for the given utterance. Informativeness indicates the ability to provide new information instead of copying from the utterance (Zhou et al., 2018a). All responses of sampled 100 case are selected from four best methods: CCM, GPT-2 (conv), ConceptFlow and Golden Response. The responses are scored from 1 to 4 by five judges.

The model performance is listed in Table 3. The human evaluation is divided into two parts: Average Score and Best@1 ratio, where Best@1 ratio indicates the fraction of judges consider the corresponding response as the best. ConceptFlow outperforms all baseline models on all scenarios. This convincing result demonstrates the advantage of explicitly modeling conversation flow with semantics: ConceptFlow outperforms GPT-2 with one-third parameters. More details of human evaluation are presented in Appendix D.

Table 3: Human Evaluation on Appropriate (Appro.) and Informativeness (Infor.). The Average Score calculates the average of human judgment score. Best@1 Ratio indicates the fraction of judges consider the case as the best. The parameter number is also presented.

Model	Parameters	Average Score		Best@1 Ratio	
		Appro.	Infor.	Appro.	Infor.
CCM	35.6M	1.802	1.802	17.0%	15.6%
GPT-2 (conv)	124M	2.100	1.992	26.2%	23.6%
ConceptFlow	35.3M	2.690	2.192	30.4%	25.6%
Golden Response	-	2.902	3.110	67.4%	81.8%



(a) Golden Concept Coverage. (b) Golden Concept Selection. (c) Response Generation.

Figure 2: Comparison of flow concept selection methods. Base only considers the central concepts. Random randomly selects two-hop concepts. Gold incorporates golden concepts in the response with random negatives. Full chooses two-hop concepts with ConceptFlow's graph attention.

4.2 Ablation Study

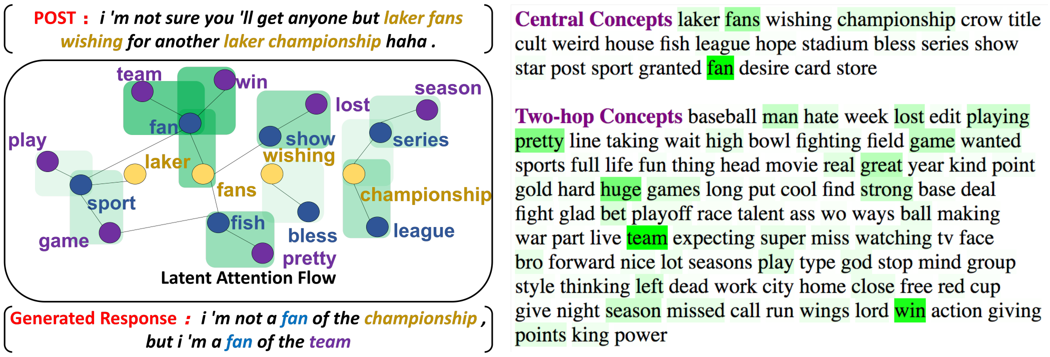
This part studies the effectiveness of the learned latent ConceptFlow. Figure 2 shows golden concept coverage, effectiveness for golden concept selection and perplexity of response generation of four different strategies to select latent concepts. Base only considers central concept graph. Random, Gold, and Full add two-hop concepts in three different ways: Random selects concepts randomly, Gold selects all golden concepts with random negatives, and Full is our method that selects by learned graph attentions.

As shown in Figure 2(a), Random has almost the same coverage with Base, while ConceptFlow (Full) performs better than Random by a large scale. This confirms the concept selection in ConceptFlow effectively selects more meaningful outer concepts for conversation generation. Then the effectiveness of two-hop concept selection strategies is presented in Figure 2(b). Full outperforms all models with Precision, Recall and F1. The ConceptFlow filters unrelated concepts and chooses underlying concepts to enhance the central graph understanding.

The high-quality latent concept flow leads to ConceptFlow's advanced performances in Figure 2(c). Interestingly, ConceptFlow even outperforms Gold in Perplexity, even Gold includes all two-hop concepts from the golden response. This shows that the "negatives" selected by ConceptFlow, even not directly appear in the target response, are also only topic and related, thus provide more meaningful information than Gold's random negatives. More results are presented in Appendix A.2.

4.3 Case Study

Figure 3 presents a case of ConceptFlow to demonstrate model effectiveness. The attention score b^{e_i} and z^g on central concepts and two-hop concepts are illustrated. The championship of zero-hop, fan of one-hop and team of two-hop receive more attention than others and are used by ConceptFlow to generate the response. On the other hand, some concepts, such as win and pretty, are filtered by the gate σ. More examples are listed in Appendix B.


POST : i 'm not sure you 'll get anyone but laker fans wishing for another laker championship haha .

Generated Response : i 'm not a fan of the championship , but i 'm a fan of the team

team
win
fan
lost
season
show
wishing
series
play
laker
sport
fans
championship
fish
bless
league
game
pretty
Latent Attention Flow
Central Concepts laker fans wishing championship crow title cult weird house fish league hope stadium bless series show star post sport granted fan desire card store
Two-hop Concepts baseball man hate week lost edit playing pretty line taking wait high bowl fighting field game wanted sports full life fun thing head movie real great year kind point gold hard huge games long put cool find strong base deal fight glad bet playoff race talent ass wo ways ball making war part live team expecting super miss watching tv face bro forward nice lot seasons play type god stop mind group style thinking left dead work city home close free red cup give night season missed call run wings lord win action giving points king power

Figure 3: Case Study (Best viewed in color). **Left:** Attention flow in commonsense concept graph, where zero-hop concepts, one-hop concepts and two-hop concepts are highlighted. **Right:** Attention scores over all concepts. Darker green indicates higher attention scores.

5 Related Work

Natural language generation (NLG) has achieved promising results with the sequence-to-sequence model (Sutskever et al., 2014) and helped build end-to-end conversation systems (Shang et al., 2015; Vinyals & Le, 2015; Li et al., 2016; Wu et al., 2019). Recently, pre-trained language models, such as ELMO (Devlin et al., 2019), BERT (Peters et al., 2018) and GPT-2 (Radford et al., 2016), further improve the NLG performance with large-scale unlabeled data. Nevertheless, the degenerating irrelevant, off-topic, and non-useful response is still one of the main challenges in conversational generation (Tang et al., 2019; Zhang et al., 2018; Gao et al., 2019).

Some work focuses on conversation generation with unstructured texts (Ghazvininejad et al., 2018; Vougiouklis et al., 2016; Xu et al., 2016; Long et al., 2017), while others extract knowledge with Convolutional Neural Network (CNN) (Long et al., 2017) or store knowledge with memory network (Ghazvininejad et al., 2018) to generate better conversation response.

The structured knowledge graphs include rich semantics about concepts and relations. Lots of previous studies focus on domain-targeted dialog system based on domain-specific knowledge base (Xu et al., 2017; Zhu et al., 2017; Gu et al., 2016). To generate the response with a large-scale commonsense knowledge base, Zhou et al. (2018a) and Liu et al. (2018) utilize graph attention and knowledge diffusion to select knowledge semantics for better user post understanding and response generation.

Different from previous research, ConceptFlow models the conversation flow explicitly with the commonsense graph and presents a novel attention mechanism using Graph Neural Network to guide the conversation flow in the latent concept spaces.

6 Conclusion

In this paper, we present ConceptFlow, which models the conversation flow explicitly as transitions in the latent concept space in order to generate more meaningful responses. Our experiments on the Reddit conversation dataset illustrate the advantages of ConceptFlow over previous conversational systems that also use prior knowledge, as well as our fine-tuned GPT-2 systems, though the latter uses much more parameters. Our studies confirm the source of this advantage mainly derive from the high quality and high coverage latent concept flow, which is effectively captured by ConceptFlow's graph attentions. Our human evaluation demonstrates that ConceptFlow generates more appropriate and informative responses by explicit modeling of the latent conversation structure.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, 2013.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

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

Hao Fang, Hao Cheng, Maarten Sap, Elizabeth Clark, Ari Holtzman, Yejin Choi, Noah A Smith, and Mari Ostendorf. Sounding board: A user-centric and content-driven social chatbot. *arXiv preprint arXiv:1804.10202*, 2018.

Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. Jointly optimizing diversity and relevance in neural response generation. *arXiv preprint arXiv:1902.11205*, 2019.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. A knowledge-grounded neural conversation model. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*, 2016.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

Alon Lavie and Abhaya Agarwal. 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*, pp. 228–231. Association for Computational Linguistics, 2007.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1192–1202, 2016.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. Knowledge diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1489–1498, 2018.

Robert Logan, Nelson F Liu, Matthew E Peters, Matt Gardner, and Sameer Singh. Baracks wife hillary: Using knowledge graphs for fact-aware language modeling. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pp. 5962–5971, 2019.

Yinong Long, Jianan Wang, Zhen Xu, Zongsheng Wang, Baoxun Wang, and Zhuoran Wang. A knowledge enhanced generative conversational service agent. In *DSTC6 Workshop*, 2017.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics, 2002.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014. URL `http://www.aclweb.org/anthology/D14-1162`.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pp. 2227–2237, 2018.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/language-unsupervised/language_understanding_paper. pdf*, 2016.

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

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*, 2015.

Robert Speer and Catherine Havasi. Representing general relational knowledge in conceptnet 5. In *LREC*, pp. 3679–3686, 2012.

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William W Cohen. Open domain question answering using early fusion of knowledge bases and text. *arXiv preprint arXiv:1809.00782*, 2018.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

Jianheng Tang, Tiancheng Zhao, Chengyan Xiong, Xiaodan Liang, Eric P Xing, and Zhiting Hu. Target-guided open-domain conversation. *arXiv preprint arXiv:1905.11553*, 2019.

Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

Pavlos Vougiouklis, Jonathon Hare, and Elena Simperl. A neural network approach for knowledge-driven response generation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 3370–3380, 2016.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*, 2019.

Chien Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, and Pascale Fung. Transferable multi-domain state generator for task-oriented dialogue systems. 2019.

Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, and Xiaolong Wang. Incorporating loose-structured knowledge into lstm with recall gate for conversation modeling. *arXiv preprint arXiv:1605.05110*, 3, 2016.

Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, and Xiaolong Wang. Incorporating loose-structured knowledge into conversation modeling via recall-gate lstm. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 3506–3513. IEEE, 2017.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems*, pp. 1810–1820, 2018.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pp. 4623–4629, 2018a.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. A dataset for document grounded conversations. *arXiv preprint arXiv:1809.07358*, 2018b.

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

A Supplementary Results

A.1 Supplementary Results for Overall Experiments

Table 4: More Metrics on Automatic Relevance Between Generated Response and Golden Response. Concept-PPL is the method used for calculating Perplexity in CCM (Zhou et al., 2018a), which combines the distribution of both words and concepts together. The Concept-PPL is meaningless when utilizing different numbers of concepts (more concepts included, better Perplexity shows).

Model	Bleu-1	Bleu-2	Bleu-3	Nist-1	Nist-2	Nist-3	Concept-PPL
Seq2Seq	0.1702	0.0579	0.0226	1.0230	1.0963	1.1056	-
MemNet	0.1741	0.0604	0.0246	1.0975	1.1847	1.1960	46.85
CopyNet	0.1589	0.0549	0.0226	0.9899	1.0664	1.0770	40.27
CCM	0.1413	0.0484	0.0192	0.8362	0.9000	0.9082	39.18
GPT-2 (lang)	0.1705	0.0486	0.0162	1.0231	1.0794	1.084	-
GPT-2 (conv)	0.1765	0.0625	0.0262	1.0734	1.1623	1.1745	-
ConceptFlow	**0.2451**	**0.1047**	**0.0493**	**1.6137**	**1.7956**	**1.8265**	26.76

Table 5: More Metrics on Repetitiveness of Generated Response. The coverage is calculated between generated response and user post, where lower means better.

Model	Bleu-1	Bleu-2	Bleu-3	Nist-1	Nist-2	Nist-3	Rouge-1
Seq2Seq	0.1855	0.0694	0.0292	1.2114	1.3169	1.3315	**0.1678**
MemNet	0.2240	0.1111	0.0648	1.6740	1.9594	2.0222	0.2216
CopyNet	0.2042	0.0991	0.056	1.5072	1.7482	1.7993	0.2104
CCM	**0.1667**	0.0741	0.0387	**1.1232**	**1.2782**	**1.3075**	0.1953
GPT-2 (lang)	0.2124	0.0908	0.0481	1.5105	1.7090	1.7410	0.1817
GPT-2 (conv)	0.2537	0.1498	0.1044	1.9562	2.4127	2.5277	0.2522
ConceptFlow	0.1850	**0.0685**	**0.0281**	1.3325	1.4600	1.4729	0.1777

A.2 Supplementary Results for Ablation Study

Table 6: Bleu and Nist for Relevance of Ablation Study. The metrics are calculated between generated responses and golden responses.

Version	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Nist-1	Nist-2	Nist-3	Nist-4
Base	0.1705	0.0577	0.0223	0.0091	0.9962	1.0632	1.0714	1.0727
Random	0.1722	0.0583	0.0226	0.0092	1.0046	1.0726	1.0810	1.0823
Gold	0.1734	0.0586	0.0230	0.0097	1.0304	1.0992	1.1081	1.1096
Full	**0.2265**	**0.0928**	**0.0417**	**0.0195**	**1.4550**	**1.6029**	**1.6266**	**1.6309**

Table 7: Rouge and Meteor for Relevance and Dist for Diversity of Ablation Study. Rouge and Meteor are calculated between generated responses and golden responses, while Dist indicates the diversity of generated responses.

Version	Rouge-1	Rouge-2	Rouge-L	Meteor	Dist-1	Dist-2
Base	0.1526	0.0192	0.1242	0.0665	0.0129	0.0542
Random	0.1537	0.0193	0.1244	0.0657	0.0134	0.0567
Gold	0.1562	0.0201	0.1256	0.0641	0.0153	0.0699
Full	**0.2101**	**0.0395**	**0.1719**	**0.0874**	**0.0203**	**0.1126**

B Cases

Table 8: Some Cases Generated by Different Models. Tokens from seed concepts, one-hop concepts and two-hop concepts are highlighted.

Post	**actually i stayed at the building right next to the lighthouse . i believe it was a boiler room . another group was already sleeping at the lighthouse .**
CCM	i 'm not sure if you 're joking , but i 'm not sure if you 're talking about the lighthouse . i 'm not sure if it was a blanket or a blanket .
GPT-2 (conv)	i 'm pretty sure it was a room with a door .
ConceptFlow	good luck , i 'm sure there are some good things in the house .
Post	**i think it 's to get their parents to care for them . predators do n't give a crap . they target young animals because they are the weakest .**
CCM	i think they are the weakest . they do n't give a shit about their children .
GPT-2 (conv)	i 'm sure they do n't care about their kids .
ConceptFlow	i have a friend who does n't care about the parents at all . it 's a great way to get a job at once a year .
Post	**give him time ? you know he hit . 162 last year right ? red sox could n't wait to get rid of him , i do n't think he will be coming out of it**
CCM	i think he 'll be fine . i think he 'll be fine . i think he 'll be a good starter for the next few seasons .
GPT-2 (conv)	i think he 'll be fine . he 's a great player , but he 's not a great player .
ConceptFlow	he 's not going to give up the money , but the media will just ignore him .
Post	**there 's an offline mode for steam , and it works fine for me .**
CCM	i 'm not sure if it works for offline mode , but i 'm not sure if it works for steam or steam . i 'm not sure if it works for online or not .
GPT-2 (conv)	i 'm not sure if it 's offline or not , but i 've never had a problem with it .
ConceptFlow	you 're right . it 's not offline . you 're going to have to wait for the next update .
Post	**at this point it may be more beneficial for him to keep playing as much as he can before the next season . trick his body into thinking he is still in the season .**
CCM	i think he 's just going to be a great player .
GPT-2 (conv)	i 'm sure he 'll be fine . he 's a great player , but he 's not a top 10 player .
ConceptFlow	he 's been playing the same game for the last two years . he 's been playing the same way .
Post	**no i 've never felt like i had friends whoops**
CCM	i 'm sorry , i feel you .
GPT-2 (conv)	i 've never felt like i had friends whoops . i 've never felt like i 've had friends whoops .
ConceptFlow	you 're gonna have to call your friend !
Post	**i 've known how to change the music , it does n't feel natural . thanks for the advice though .**
CCM	i 'm not sure if you 're joking , but i 'm not sure if you 're being sarcastic or not . i 'm not sure if you 're being sarcastic or not , but i 'm not sure if you 're being sarcastic or not . i 'm pretty sure it 's just a video .
GPT-2 (conv)	i 've been playing with a lot of different music and it 's always been the same .
ConceptFlow	i did n't even notice the natural sound . i 'll check it out !
Post	**are milan and inter no longer top clubs given that they currently do n't have the finances to compete with the best ?**
CCM	milan is n't a top club , but they are n't top clubs . they are n't the only ones who are n't in the world .
GPT-2 (conv)	i think they are . they are a club that has been in the top 4 for a long time .
ConceptFlow	the best part of the team is the way they are , and we do n't want to play the best team in the country .

C concept Selection

With the concept graph growing, more concepts are considered and a concept selector is needed for concept filter. The concept selector aims to select related two-hop concepts based on the sum of attention scores for each time t over entire two-hop concepts:

$$\alpha_n = \sum_{t=1}^{n} \text{softmax}(\vec{s}_t \cdot \vec{e}_k), \tag{20}$$

where $\vec{s}_t$ is the t-th time **decoder** output representation and $\vec{e}_k$ denotes the concept e_k's embedding. Then top-k concepts are reserved to construct the two-hop concept graph G^2 with central concept graph. Moreover, central concepts are all reversed because of the high correlation to the conversation topic and acceptable computation complexity.

D Agreement of Human Evaluation

Table 9: Fleiss' Kappa for Human Evaluation.

Model	Appropriateness	Informativeness
ConceptFlow-CCM	0.3724	0.2641
ConceptFlow-GPT2	0.2468	0.2824

For human evaluation, 100 cases with four responses from CCM, GPT-2 (conv), ConceptFlow and Golden Response are sampled and listed in an Excel file with randomly sort. A group of human judges are asked to score each response with 1 to 4 based on the quality of appropriateness and informativeness respectively, without knowing any clues of the source of response, thus the impartiality and objectivity of the evaluation can be guaranteed.

To further demonstrate the consistency among human judges, the agreement of human evaluation for CCM, GPT-2 (conv) and ConceptFlow are presented in Table 9. For each case, the result scores from two baseline models is compared with ConceptFlow and is divided into three categories: win, tie and loss. Then human evaluation agreement is indicated by Fleiss' Kappa. All agreement values fall into the fair level of agreement, which confirms the quality of human evaluation.

E Data Study

Table 10: Statistics of Coverage and Number of Zero-hop, One-hop, Two-hop and Three-hop Concept Graph.

Concept	Concept Number	Coverage Ratio	Coverage Number
Zero-hop	5.8	9.81%	0.579
+ One-hop	98.6	38.78%	2.292
+ Two-hop	880.8	61.37%	3.627
+ Three-hop	3769.1	81.58%	4.821

To determine the gown deep of concept graph for conversation generation, some statistics are presented in Table 10. The two-hop deep concept graph covers more than 61% golden concepts appearing in the response with acceptable computational efficiency. With growing to the three-hop, the number of concepts is increased dramatically with only about one extra golden concept for each case, thus the outer concept ends in two-hop concepts because of the close connection with the topic and the endurable computation complexity.