Language Generation with Multi-Hop Reasoning on Commonsense Knowledge Graph

Haozhe Ji[1], Pei Ke[1], Shaohan Huang[2], Furu Wei[2], Xiaoyan Zhu[1], Minlie Huang[1]*

[1]Department of Computer Science and Technology, Institute for Artificial Intelligence,
State Key Lab of Intelligent Technology and Systems,
Beijing National Research Center for Information Science and Technology,
Tsinghua University, Beijing 100084, China
[2]Microsoft Research

{jhz20,kp17}@mails.tsinghua.edu.cn, {shaohanh, fuwei}@microsoft.com,
{zxy-dcs,aihuang}@tsinghua.edu.cn

Abstract

Despite the success of generative pre-trained language models on a series of text generation tasks, they still suffer in cases where reasoning over underlying commonsense knowledge is required during generation. Existing approaches that integrate commonsense knowledge into generative pre-trained language models simply transfer relational knowledge by post-training on individual knowledge triples while ignoring rich connections within the knowledge graph. We argue that exploiting both the structural and semantic information of the knowledge graph facilitates commonsense-aware text generation. In this paper, we propose *Generation with Multi-Hop Reasoning Flow* (GRF) that enables pre-trained models with dynamic multi-hop reasoning on multi-relational paths extracted from the external commonsense knowledge graph. We empirically show that our model outperforms existing baselines on three text generation tasks that require reasoning over commonsense knowledge. We also demonstrate the effectiveness of the dynamic multi-hop reasoning module with reasoning paths inferred by the model that provide rationale to the generation.[1]

1 Introduction

Despite the recent success of pre-trained language models such as GPT-2 (Radford et al., 2019) on various language generation tasks, these models are still struggling on generation tasks that require reasoning over commonsense knowledge that is not *explicitly* stated in the context. For example, Figure 1 illustrates an example in the story ending generation task, where external commonsense knowledge in the form of relational paths can guide the generation of the key concepts "substance" and

---

* Corresponding author
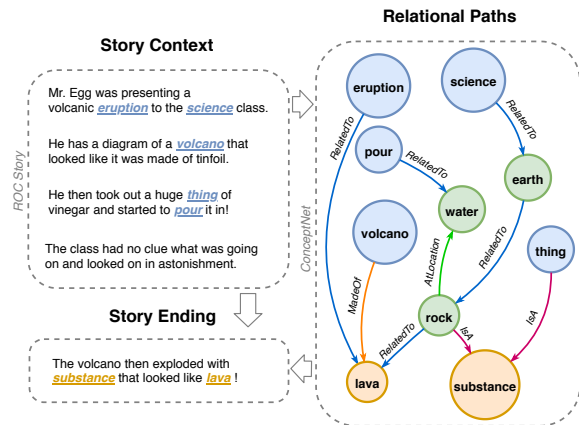[1]The source code is available at https://github.com/cdjhz/multigen.

Figure 1: An example of using structural relational knowledge as commonsense grounding in story ending generation. Blue nodes correspond to the concepts in the context, orange nodes correspond to those in the story ending and gree nodes are intermediate concepts that connect the evidence chain.

"lava" in the story ending by providing background knowledge such as (volcano,MadeOf,lava) besides the story context. Although pre-trained models have been demonstrated to possess commonsense reasoning ability (Trinh and Le, 2018) by *implicitly* learning some relational patterns from large-scale corpora, they do not fully utilize the commonsense knowledge bases that provide more explicit knowledge grounding.

To address this defect, incorporating external commonsense knowledge to enhance models' reasoning ability has been widely explored (Lin et al., 2019; Ye et al., 2019; Lv et al., 2019). In language generation, previous work (Bhagavatula et al., 2020; Guan et al., 2020) transfers commonsense knowledge into pre-trained language models by utilizing triple information in commonsense knowledge bases such as ConceptNet (Speer and Havasi, 2012) and ATOMIC (Sap et al., 2019).

However, this approach has two drawbacks.

First, recovering knowledge triples at the post-training stage (Guan et al., 2020) hardly enables the model to utilize the encoded knowledge in fine-tuning generation tasks which requires reasoning over underlying commonsense knowledge. Second, it ignores the abundant structural relational relevance of the concepts in the knowledge graph (Guan et al., 2020; Bhagavatula et al., 2020) that may provide multiple plausible evidence for complex reasoning. Thus a richer and more explicit way of utilizing external commonsense knowledge is to exploit both *structural* and *semantic* information of the knowledge graph and reason over multi-hop relational paths where multiple connected triples provide chains of evidence for grounded text generation.

In this paper, we propose *Generation with Multi-Hop Reasoning Flow* (GRF), a generation model that performs multi-hop reasoning on the external knowledge graph for knowledge-enriched language generation. The model operates on the sub-graph extended from the concepts in the input text as commonsense knowledge grounding. It first encodes the multi-relational graph with compositional operation to obtain graph-aware representations for the concepts and the relations (§3.2.1). Then, the multi-hop reasoning module performs dynamic reasoning via aggregating triple evidence along multiple relational paths to generate the salient concept under the context (§3.2.3). Finally, the generation distribution combines the probability of copying concepts from the knowledge graph and that of choosing a word from the standard vocabulary with a gate control (§3.2.4). The overall model architecture is shown in Figure 2. We conduct experiments on three commonsense-aware text generation tasks including story ending generation (Mostafazadeh et al., 2016), abductive natural language generation (Bhagavatula et al., 2020), and explanation generation for sense making (Wang et al., 2019). Results show that our model outperforms strong baselines on these tasks, thereby demonstrating the benefit of multi-hop commonsense reasoning in language generation.

Our contributions can be summarized as follows: 1) We propose GRF, a novel generation model that utilizes external structural commonsense knowledge to facilitate explicit commonsense reasoning in text generation. 2) We propose the dynamic multi-hop reasoning module that aggregates evidence along relational paths for grounded generation of some critical concepts. 3) We conduct extensive experiments including automatic and human evaluation on three commonsense-aware text generation tasks and show that our model outperforms various selective baselines. We also visualize reasoning paths inferred by the model to demonstrate the effectiveness of the multi-hop reasoning module.

2 Related Work

2.1 Commonsense-Aware Neural Text Generation

Incorporating commonsense knowledge is essential for text generation to augment the limited textual information. In dialogue generation, Zhou et al. (2018) enriched the context representations of the post with neighbouring concepts on ConceptNet using graph attention. In story ending generation, Guan et al. (2019) proposed incremental encoding with multi-source attention to incorporate one-hop knowledge graph for concepts in the story context. In topic-to-essay generation, Yang et al. (2019) augmented the generator with a concept memory that updated dynamically with gate mechanism. Recently, some work also attempted to integrate external commonsense knowledge into generative pre-trained language models such as GPT-2 (Radford et al., 2019). Guan et al. (2020) conducted post-training on sythetic data constructed from commonsense knowledge bases by translating triplets into natural language texts using templates. Bhagavatula et al. (2020) transferred embeddings of COMeT (Bosselut et al., 2019), a GPT-2 model fine-tuned to generate the tail entity of a triple in commonsense knowledge graph, into another GPT-2 model for text generation. In comparison, our model utilizes both structural and semantic information of the commonsense knowledge graph during generation and does not suffers from the catastrophic forgetting problem (Kirkpatrick et al., 2016) caused by implicit knowledge transferring.

2.2 Multi-Hop Reasoning on Graph Structure

Performing explicit multi-hop reasoning on graph structure has been demonstrated to be an effective approach for query answering over incomplete knowledge graphs (Das et al., 2018; Chen et al., 2018; Lin et al., 2018), multi-hop question answering (Bauer et al., 2018; Cao et al., 2019; Qiu et al., 2019) and dialogue generation (Tuan et al.,
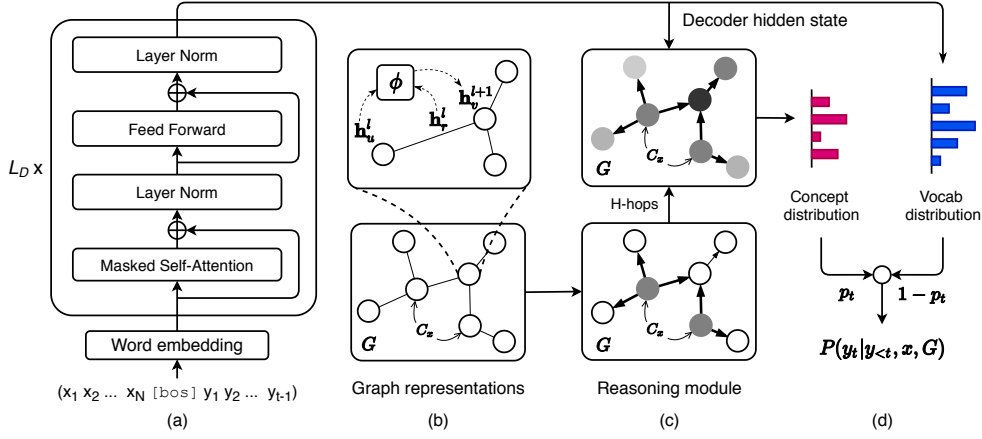
Figure 2: Model architecture. (a) Context modeling with pre-trained transformer (§3.2.2). (b) The model encodes the multi-relational graph with non-parametric operation $\phi(\cdot)$ to combine relations and concepts (§3.2.1). (c) The multi-hop reasoning module aggregates evidence from source concepts $C_{\boldsymbol{x}}$ along structural paths to all nodes where shade indicates the node score (§3.2.3). (d) The final generation distribution with gate control (§3.2.4).

2019; Moon et al., 2019; Liu et al., 2019). Particularly, reasoning on knowledge graphs to answer relational query typically adopts REINFORCE to learn concrete policies to search for entities or relations. In multi-hop question answering tasks, the reasoning process is augmented with entity graph (Cao et al., 2019; Qiu et al., 2019) or concept paths (Bauer et al., 2018) to enhance semantic connections among document segments. In dialogue generation, Tuan et al. (2019) modeled multiple hops on relationship graphs with a Markov transition matrix. Liu et al. (2019) proposed a two-stage architecture that selected information from a knowledge graph for further generating the response. Compared with these generation models that operate on knowledge graphs within a specific domain, our focus is to utilize general commonsense knowledge to supply evidence for text generation.

3 Methodology

3.1 Problem Formulation

In this paper, we focus on text generation tasks where reasoning over external commonsense knowledge is required. Without loss of generality, the input source is a text sequence $\boldsymbol{x} = (x_1, x_2, \cdots, x_N)$ which may consist of several sentences. The output target is another text sequence $\boldsymbol{y} = (y_1, y_2, \cdots, y_M)$. To facilitate the reasoning process, we resort to an external commonsense knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ denotes the concept set and $\mathcal{E}$ denotes the relations connecting these concepts. Since direct reasoning on the com-

plete graph is intractable, we extract a sub-graph $G = (V, E)$ given the input text where $V \subset \mathcal{V}$ and $E \subset \mathcal{E}$. The sub-graph consists of inter-connected H-hop paths starting from the source concepts $C_{\boldsymbol{x}}$ extracted from the input text. We only consider concepts with 1-gram surface texts. The task is then formulated as generating the best hypothesis $\hat{\boldsymbol{y}}$ which maximizes the following conditional probability:

$$\hat{\boldsymbol{y}} = \mathrm{argmax}_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{x}, G). \quad (1)$$

We leave the detailed sub-graph extraction process in §4.2 and describe our proposed model in the next section.

3.2 Generation with Multi-Hop Reasoning Flow

3.2.1 Static Multi-Relational Graph Encoding

Graph Neural Network (GNN) frameworks, such as graph convolution network (GCN) (Kipf and Welling, 2017) and graph attention network (GAT) (Velickovic et al., 2018), have been shown effective at encoding graph-structured data by aggregating node information from local neighbours. To model the relational information in the knowledge graph, R-GCN (Schlichtkrull et al., 2018) generalizes GCN with relation-specific weight matrices but is reported to be over-parameterized (Marcheggiani and Titov, 2017; Schlichtkrull et al., 2018). We follow Vashishth et al. (2020) and use a non-parametric compositional operation $\phi(\cdot)$ to combine the node embedding and the relation embedding. Specifically,

given the input graph $G = (V, E)$ and a GCN with L_G layer, for each node $v \in V$ we update the node embedding at the $l + 1$-th layer by aggregating information from its local neighbours $\mathcal{N}(v)$ which consist of pairs of node u and the connected relation r.

$$\boldsymbol{o}_v^l = \frac{1}{|\mathcal{N}(v)|} \sum_{(u,r) \in \mathcal{N}(v)} \mathbf{W}_N^l \phi(\boldsymbol{h}_u^l, \boldsymbol{h}_r^l), \quad (2)$$

$$\boldsymbol{h}_v^{l+1} = \text{ReLU}\Big(\boldsymbol{o}_v^l + \mathbf{W}_S^l \boldsymbol{h}_v^l\Big), \quad (3)$$

where $\boldsymbol{h}_v^0$ is initialized by looking up the word embedding and $\boldsymbol{h}_r^0$ by the relation-type embedding. $\mathbf{W}_N^l$ and $\mathbf{W}_S^l$ are two learnable weight matrices specific to the l-th layer. We define the compositional operation as $\phi(\boldsymbol{h}_u, \boldsymbol{h}_r) = \boldsymbol{h}_u - \boldsymbol{h}_r$ inspired by the TransE model (Bordes et al., 2013).

The relation embedding is also updated via another linear transformation.

$$\boldsymbol{h}_r^{l+1} = \mathbf{W}_R^l \boldsymbol{h}_r^l. \quad (4)$$

Finally, we obtain node embeddings $\boldsymbol{h}_v^{L_G}$ and relation embeddings $\boldsymbol{h}_r^{L_G}$ that encode the static graph context for dynamic reasoning during decoding.

3.2.2 Context Modeling with Pre-Trained Transformer

We adopt the GPT-2 model (Radford et al., 2019), a pre-trained multi-layer transformer decoder to model the contextual dependency of the text sequence. The input to the model is the concatenation of the source and the target sequence: $\boldsymbol{s} = (x_1, \cdots, x_N, [\text{bos}], y_1, \cdots, y_M)$.

$$\boldsymbol{h}_t^0 = \boldsymbol{e}_t + \boldsymbol{p}_t, \quad (5)$$

$$\boldsymbol{h}_t^l = \text{T\_block}(\mathbf{H}_{\leq t}^{l-1}), l \in [1, L_D] \quad (6)$$

$$P(s_t | \boldsymbol{s}_{<t}) = \text{softmax}(\mathbf{W}_{LM} \boldsymbol{h}_t^{L_D} + \boldsymbol{b}) \quad (7)$$

where $\boldsymbol{e}_t$ and $\boldsymbol{p}_t$ are the token embedding vector and the positional embedding vector. T_block is the transformer block with masked self-attention. The final hidden state at the t-th time step $\boldsymbol{h}_t^{L_D}$ which encodes the context information is used as the input to the multi-hop reasoning module.

3.2.3 Dynamic Multi-Hop Reasoning Flow

To perform explicit reasoning on the graph structure during generation, we devise a dynamic reasoning module that utilizes both structural patterns of the knowledge graph and contextual information to propagate evidence along relational paths at each decoding step.

Specifically, the module broadcasts information on G by updating the score of outer nodes with their visited neighbours for multiple hops until all the nodes on G are visited. Initially, nodes correspond to the concepts in $C_{\boldsymbol{x}}$ are given a score of 1 while other unvisited nodes are assigned with 0.

For the unvisited node $v \in V$, its node score $ns(v)$ is computed by aggregating evidence from $\mathcal{N}_{in}(v)$ which denotes the set of visited node u and its edge r that directly connects v.

$$ns(v) = \underset{(u,r) \in \mathcal{N}_{in}(v)}{f} \Big(\gamma \cdot ns(u) + R(u, r, v)\Big), \quad (8)$$

where γ is a discount factor that controls the intensity of the information flow from the previous hops. $f(\cdot)$ is the aggregator that assembles scores from connected nodes. We consider two forms of aggregators: $\max(\cdot)$ and $\text{mean}(\cdot)$. We use $\max(\cdot)$ for the main results and present the results with $\text{mean}(\cdot)$ in the ablation study.

$R(u, r, v)$ is the triple relevance that reflects the relevancy of the evidence given by the triplet (u, r, v) under the current context. We compute the triple relevance as follows:

$$R(u, r, v) = \sigma(\boldsymbol{h}_{u,r,v}^{\text{T}} \mathbf{W}_{sim} \boldsymbol{h}_t^{L_D}), \quad (9)$$

$$\boldsymbol{h}_{u,r,v} = [\boldsymbol{h}_u^{L_G}; \boldsymbol{h}_r^{L_G}; \boldsymbol{h}_v^{L_G}]. \quad (10)$$

After H hops, the final distribution over the nodes is obtained by a normalization.

$$P(c_t | \boldsymbol{s}_{<t}, G) = \text{softmax}_{v \in V}(ns(v)), \quad (11)$$

where c_t is the concept of the selected node at the t-th time step.

Intuitively, the reasoning module learns to dynamically distribute along the paths by considering the triple evidence according to the current decoder state.

3.2.4 Generation Distribution with Gate Control

The final generation distribution combines the distribution over the concepts (Eq. 11) and the distribution over the standard vocabulary (Eq. 7). We use a soft gate probability g_t which denotes whether to copy a concept in the generation to control the weight of the two distributions similar to the copy mechanism (Gu et al., 2016; See et al., 2017).

$$g_t = \sigma\Big(\mathbf{W}_{gate} \boldsymbol{h}_t^{L_D}\Big). \quad (12)$$

The final output distribution is the linear combination of the two distributions weighted by g_t and $1 - g_t$ respectively.

$$P(y_t|\boldsymbol{y}_{<t}, \boldsymbol{x}, G) = g_{t+N} \cdot P(c_{t+N}|\boldsymbol{s}_{<t+N}, G)$$
$$+ (1 - g_{t+N}) \cdot P(s_{t+N}|\boldsymbol{s}_{<t+N}), \quad (13)$$

where N is the length of the input text sequence.

3.3 Training and Inference

To train the proposed model, we minimize the negative log-likelihood of generating the ground truth target sequence $\boldsymbol{y}^{\text{gold}} = (y_1, y_2 \cdots, y_M, [\text{eos}])$.

$$\mathcal{L}_{gen} = \sum_{t=1}^{M+1} - \log P(y_t^{\text{gold}}|\boldsymbol{y}_{<t}^{\text{gold}}, \boldsymbol{x}, G). \quad (14)$$

We add an auxiliary gate loss $\mathcal{L}_{gate}$ to supervise the probability of selecting a concept or a generic word. We additionally introduce a weak supervision $\mathcal{L}_{weak}$ to induce the predicted triple relevances to match the heuristic labels of edges obtained by breadth-first search from the source concepts to the concepts in $\boldsymbol{y}^{\text{gold}}$ on the graph. Both loss functions take the form of binary cross-entropy. We observe that both loss terms encourage the model to learn multi-hop reasoning on the graph more effectively.

The final loss to be optimized is the linear combination $\mathcal{L}_{gen} + \alpha\mathcal{L}_{gate} + \beta\mathcal{L}_{weak}$.

During the inference stage, the input to the model is $(x_1, \cdots, x_N, [\text{bos}])$. The model generates a token at a time and concatenates it to the input sequence to generate the next token. The generation process terminates when the special ending symbol $[\text{eos}]$ is generated.

4 Experiments

4.1 Datasets and Metrics

The statistics of the datasets are shown in Table 1. **Story Ending Generation** (SEG) is to generate a reasonable ending given a four-sentence story context. The stories come from ROCStories corpus (Mostafazadeh et al., 2016). We use the same data split as Guan et al. (2019).
Abductive NLG (αNLG) is to generate an explanatory hypothesis given two observations: O_1 as the cause and O_2 as the consequence. We use the official data split[2] from Bhagavatula et al. (2020).

---

[2] https://github.com/allenai/abductive-commonsense-reasoning

Explanation Generation (EG) is to generate an explanation given a counter-factual statement for sense-making (Wang et al., 2019). We randomly split 85% of the data as the training set, 10% as the test set, and the latter as the development set.

For automatic evaluation, we use metrics including BLEU-4 (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), ROUGE-L (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) to evaluate the abductive NLG and the explanation generation tasks. We follow common practice in story generation (Guan et al., 2019, 2020) and use BLEU-1/2 to evaluate the generated endings. We also adopt Distinct-n (Li et al., 2016) to measure the diversity of the generated endings.

4.2 Extracting Sub-Graphs as Knowledge Grounding

To supply knowledge grounding for language generation, we use ConceptNet (Speer and Havasi, 2012) as the commonsense knowledge base. Each triple (h, r, t) in ConceptNet denotes that the head concept h has a relation r with the tail concept t. To condense the knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ we group the original 42 relation types into 17 following Lin et al. (2019) and add reversed links (t, r^{-1}, h) to the graph (Lin et al., 2018; Das et al., 2018).

We extract a sub-graph $G = (V, E)$ from $\mathcal{G}$ which consists of multiple inter-connected paths starting from the source concepts $C_{\boldsymbol{x}}$ in the input sequence. To recognize concepts from the input text sequence, we perform fuzzy matching with the lemmatized form of the surface texts using Spacy[3] and filter out stop words. Following Guan et al. (2019), we only consider verbs and nouns as our candidate concepts since we find the extracted graph is much noisier with all the matched concepts.

Specifically, we iterate the following process for H hops: starting from the nodes in the current sub-graph (initialized by $C_{\boldsymbol{x}}$) and search for the direct neighbours of each node and preserve top-B nodes with the connected edges to enlarge the sub-graph. For each candidate node, the selection is based on its incoming degree of this node. The incoming degree of a candidate node v is defined as the number of nodes in the current sub-graph that directly connect v. Intuitively, we keep those salient concepts that are commonly visited nodes and support information flow on the graph.

---

[3] https://spacy.io/

729

Tasks	Train	Dev	Test
SEG*	90,000	4,081	4,081
αNLG	50,481	7,252	14,313
EG*	25,596	1,428	2,976

Table 1: Statistics of the datasets used in this paper. *:Examples with multiple references are counted separately.

4.3 Implementation Details

Graph statistics	EG	αNLG	SEG
Avg. # Concepts	193.1	201.6	208.5
Avg. # Triples	1094.3	1324.6	1148.6

Table 2: Statistics of the extracted subgraphs on the training sets of three datasets, including the average number of concepts and triples for each subgraph.

Our model employs the small version of GPT-2 model[4] with 12 layers, 768-dimensional hidden states, and 12 attention heads for contextual modeling and a 2-layer GCN model. We choose the $\max(\cdot)$ aggregator for the main results since it yields better performance. During sub-graph extraction, we set the maximum number of hops $H = 2$ and preserve top-$B = 100$ nodes per hop. We find this setting balances the coverage and noise of the knowledge graph. Detailed statistics of the extracted sub-graphs are presented in Table 2. To train the model, we use the Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 1 \times 10^{-8}$ and linearly decrease the learning rate to zero with no warmup. We search the best hyper-parameters according to BLEU-4 on the development set of each task. At the inference stage, we adopt beam search decoding with a beam size of 3 for our model and all the baselines we produce. We conduct all the experiments using the PyTorch framework (Paszke et al., 2017).

4.4 Compared Baselines

We produce the following baselines on three generation tasks to compare with our model:

Seq2Seq is a sequence-to-sequence model based on gated recurrent unit (GRU) with attention mechanism. We also utilize the copying mechanism (Gu et al., 2016) for the model to generate out-of-vocabulary words.

---

[4] https://github.com/huggingface/transformers

GPT2-FT is a GPT-2 model fine-tuned on the task-specific dataset with its model initialization from Radford et al. (2019).

GPT2-OMCS-FT is a commonsense-enhanced GPT-2 model first post-trained on the Open Mind Common Sense (OMCS) corpus[5] from which the ConceptNet is constructed. The model is then fine-tuned on the task-specific dataset.

We also compare our model with baseline models designated to each specific task. For story ending generation, we compare to **IE+GA** which is based on incremental encoding and graph attention (Guan et al., 2019) and **WriterForcing** that forces the attention to focus on important keyphrases and avoid generating generic words.

For abductive NLG, we compare with two baselines introduced by Bhagavatula et al. (2020): **COMeT-Txt-GPT2** which uses the output texts generated by COMeT as prefix texts to the GPT-2 model while fine-tuning, and **COMeT-Emb-GPT2** which integrates the embeddings of the outputs generated by COMeT into the GPT-2 model during fine-tuning.

4.5 Automatic Evaluation

We present the automatic evaluation results on the test sets of the explanation generation and the abductive NLG tasks in Table 3. We have the following observations:

I. Our model outperforms all the baselines that utilize pre-trained language models or incorporate external commonsense knowledge in terms of all evaluation metrics indicating that incorporating rich structural information of commonsense knowledge graphs can enhance the overall generation quality.

II. Simply post-training on commonsense knowledge source degrades the performance on these two tasks. This is possibly due to the fact that the triple-level post-trained corpus cannot provide rich semantics for the model to generalize on tasks that emphasize reasoning and explaining.

For story ending generation, we present the evaluation results in Table 4. Our model outperforms all the baselines in BLEU and distinct metrics. We also observe that post-training on external commonsense data improves the generation diversity of the pre-trained language model, which accords with the findings of Guan et al. (2020). We suspect that post-training on the commonsense data enables

---

[5] http://openmind.media.mit.edu

Models	EG				αNLG			
	BLEU-4	METEOR	ROUGE-L	CIDEr	BLEU-4	METEOR	ROUGE-L	CIDEr
Seq2Seq	6.09	24.94	26.37	32.37	2.37	14.76	22.03	29.09
COMeT-Txt-GPT2	N/A	N/A	N/A	N/A	2.73†	18.32†	24.39†	32.78†
COMeT-Emb-GPT2	N/A	N/A	N/A	N/A	3.66†	19.53†	24.92†	32.67†
GPT2-FT	15.63	38.76	37.32	77.09	9.80	25.82	32.90	57.52
GPT2-OMCS-FT	15.55	38.28	37.53	75.60	9.62	25.83	32.88	57.50
GRF	**17.19**	**39.15**	**38.10**	**81.71**	**11.62**	**27.76**	**34.62**	**63.76**

Table 3: Automatic evaluation results on the test set of EG and αNLG. Entries with N/A mean the baseline is not designated for this task. $\dagger$: we use the generation results from Bhagavatula et al. (2020).

Models	BLEU-1/2	Distinct-2/3
Seq2Seq	19.1 / 5.5	0.181 / 0.360
IE+GA	20.8 / 6.4	0.140 / 0.280
WriterForcing	16.5 / 3.7	0.335 / 0.584
GPT2-FT	25.5 / 10.2	0.304 / 0.505
GPT2-OMCS-FT	25.5 / 10.4	0.352 / 0.589
GRF	**26.1 / 11.0**	**0.378 / 0.622**

Table 4: Automatic evaluation on the test set of SEG.

Models	BLEU-4	ROUGE-L
GRF	11.62	34.62
w/ mean($\cdot$) aggregator	11.32	34.46
w/o DMRF	10.67	33.75
w/o SMGE	11.10	34.36

Table 5: Ablation study on the test set of αNLG. SMGE denotes static multi-relational graph encoding (see §3.2.1) and DMRF denotes dynamic multi-hop reasoning flow (see §3.2.3).

the model to generate concepts related to the story context, which improves the text diversity.

4.6 Human Evaluation

To evaluate the fluency and the reasonability of the generated texts under the specific task settings, we conduct pair-wise comparison with **COMeT-Emb-GPT2** on αNLG, **IE+GA** on SEG, and with two fine-tuned GPT-2 models on all the three tasks. For human evaluation, we randomly sample 100 sentences from the test set for each pair of models and obtain 1,100 sentences from five models. We recruit three annotators to make a preference among *win*, *tie* and *lose* given the input context and two outputs generated by our model and a baseline respectively according to two criteria: fluency and reasonability.

For fluency, we require the annotators to focus only on the grammatical correctness and readability of the generated results disregarding the input

context. When evaluating reasonability, the annotators are required to assess whether the generated sentence is reasonable under the given context in each task. In SEG and αNLG, annotators are asked to focus on evaluating the causal and temporal relevance of the generated results and the contexts. While on EG, annotators are mainly asked to check whether the generated results explain the counterfactual points in the statements properly.

The human evaluation results are presented in Table 6 where our model significantly outperforms compared baselines in terms of both criteria on all the datasets. Specifically, incorporating structural commonsense knowledge yields significant improvement in generating reasonable texts given the context. Table 7 shows the inter-rater agreement where five out of six annotations show moderate ($0.4 \le \kappa < 0.6$) or good agreement ($0.6 \le \kappa < 0.8$). We check the annotation results and find that the GPT-2 baselines also generate story endings with good grammar, which makes it hard for the annotators to reach a high consensus when evaluating the fluency criterion of the story ending generation task ($\kappa = 0.315$).

4.7 Ablation Study

We conduct ablation study to verify the effect of different model components. As shown in Table 5, all the components contribute to the final performance. Removing the dynamic reasoning module (w/o DMRF) results in the largest performance drop, thereby indicating that dynamic multi-hop reasoning plays a major role in this task. Ablating the graph representation module (w/o SMGE) also degrades the performance since it encodes the graph structure with relational information that benefits concept selection. We also show the results of our reasoning module with mean($\cdot$) aggregator and observe some performance drop comparing with max($\cdot$).

Models	EG				αNLG				SEG			
	Fluency		Reasonability		Fluency		Reasonability		Fluency		Reasonability	
	W	L	W	L	W	L	W	L	W	L	W	L
vs. IE+GA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.62**	0.07	0.72**	0.11
vs. COMeT-Emb-GPT2	N/A	N/A	N/A	N/A	0.31**	0.14	0.55**	0.25**	N/A	N/A	N/A	N/A
vs. GPT2-FT	0.24**	0.09	0.54**	0.21	0.15*	0.10	0.56**	0.20	0.21**	0.12	0.45**	0.19
vs. GPT2-OMCS-FT	0.18**	0.09	0.58**	0.18	0.12	0.09	0.50**	0.20	0.17*	0.11	0.40**	0.15

Table 6: Human evaluation results on three datasets. Scores indicate the percentage of *Win* (**W**) and *Lose* (**L**) when comparing our model with a baseline in terms of *fluency* and *reasonability*. Scores marked with * mean p-value < 0.05 and ** indicates p-value < 0.01 in sign test. Entries with N/A mean the baseline is not designated for this task.

Criteria	EG	αNLG	SEG
Fluency	0.615	0.543	0.315
Reasonability	0.551	0.677	0.595

Table 7: Annotator agreement. Scores denotes Fleiss' kappa (Fleiss, 1971) which evaluates the agreement from multiple annotators in terms of *fluency* and *reasonability*.
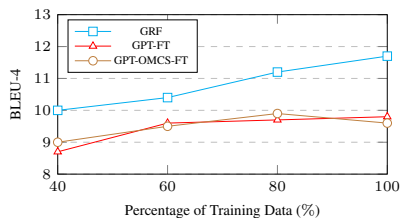

Figure 3: Performance with different amount of training data on the test set of αNLG.

Figure 4: Effect of γ in DMRF. Performance with different value of discount factor γ on the test set of αNLG.

4.8 Impact of the Size of Training Data

To demonstrate the complementary performance gain of utilizing relational paths besides textual modeling, we sample different fractions of training data of αNLG for training and evaluate on the original test set. We compare our method with knowledge-agnostic finetuning of the GPT-2 model and the commonsense-enhanced GPT-2 post-trained on OMCS. As shown in Figure 3, our model achieves consistent performance gains against the chosen baselines with different amount of training data, which demonstrates the generalization ability of the proposed model with the aid of structural relation knowledge.

4.9 Effectiveness of Dynamic Multi-Hop Reasoning

We demonstrate the effectiveness of the multi-hop reasoning module through both quantitative and qualitative analysis.

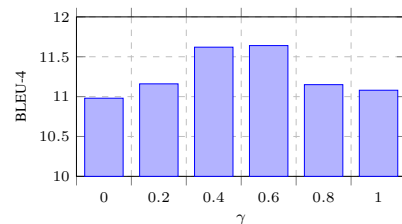We investigate the impact of the hyper-parameter γ that controls the information flow in the multi-hop reasoning module. As shown in Figure 4, the maximum performance is obtained when γ is around 0.4 and 0.6. When $\gamma \to 0$, the multi-hop reasoning module reduces to local scoring of each concept and ignores evidence accumulated on the paths. While $\gamma \to 1$, the node score increases monotonically along the paths which also hinders the model's ability to select the correct concept. Therefore, we set $\gamma = 0.5$ for all the main results of our model.

To qualitatively assess the ability of the dynamic reasoning module, we visualize a test case on αNLG with top-ranked concepts and scored reasoning paths. As shown in Figure 5, at the first hop the reasoning module starts from the source concepts "*adopt*" and "*puppy*" and assigns higher scores to neighbouring concepts which are verbs considering the generated context. At the second hop the module utilizes more plausible evidences along 2-hop reasoning paths and selects "*play*" ($g_t = 0.64$) which is more reasonable given both the observations.

4.10 Case Study

We provide some test cases on three datasets in Table 8 and observe that: **I.** Baseline models tend to generate general cases while the GRF is able to generate more specific concepts by exploring the plau-
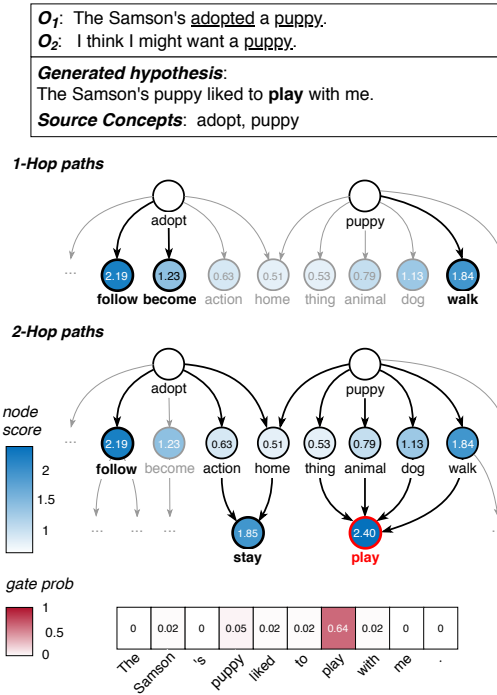
Figure 5: Visualization of a test case with inferred reasoning paths by our model. We highlight top-3 concepts with reasoning paths at **1-Hop** and **2-Hop** reasoning step respectively.

sible relations between concepts. For example in the first case, the GRF generates "expensive" which is the antonym of "cheap" under the story context. **II.** Baseline models sometimes fail to identify the transition of the narrative as shown in the third case where the GRF generates "seasick" as a plausible explanation for the transition from "cruise" to "beach". **III.** The GRF generates proper attributes of the source concepts in the input context with the aid of external commonsense knowledge as shown in the last two cases of explanation generation.

5 Conclusion

We present Generation with Multi-Hop Reasoning Flow that reasons over structured commonsense knowledge during text generation. The proposed method leverages both the structural and semantic information of the external knowledge base by performing dynamic multi-hop reasoning on the relational paths. We conduct extensive experiments and empirically show that our method outperforms existing approaches that integrate commonsense knowledge to pre-trained language models on three text generation tasks. We also demonstrate the interpretability of our method with inferred reasoning paths that provide rationale to the generated results.

Story Ending Generation	
Story Context	*I wanted a simple bike for commuting.* *So I bought a **cheap** one on sale.* *But it didn't fit me properly.* *And it was uncomfortable to ride.*
IE+GA	So I decided to buy a new one.
GPT2-FT	So I decided to buy a new bike.
GPT2-OMCS-FT	So I decided to buy a bike from a bike shop instead.
GRF	So I decided to get a more **expensive** bike.
Story Context	*Ava made shakes for her kids on a **hot** summer day.* *She called them in from play, but they dallied.* *By time they came in, the shakes were almost **melted.*** *Ava blended in more **ice** cubes and refreshed them.*
IE+GA	Then she went home and ate them.
GPT2-FT	Ava was proud of her kids for being so good at cooking.
GPT2-OMCS-FT	She was proud of her kids for being so thoughtful!
GRF	Her kids thanked her profusely for helping them **cool** off.
Abductive NLG	
Observation 1	*The Smith family went on a **cruise** for their summer vacation.*
Observation 2	*From then on, the Smiths went to the **beach** each summer instead.*
GPT2-FT	The Smith family had a great time on the beach.
GPT2-OMCS-FT	The Smith family went to the beach.
COMeT-Emb-GPT2	They didn't have a nice vacation.
GRF	The Smith family got **seasick** on the cruise.
Observation 1	*Nancy bought her dog a squeaky stuffed animal.*
Observation 2	*The dog had **ripped** the toy to shreds.*
GPT2-FT	Nancy found a toy that looked like a toy.
GPT2-OMCS-FT	Nancy found a toy that looked like a toy.
COMeT-Emb-GPT2	The squeaky stuffed animal was the first to come in.
GRF	Nancy's dog **scratched** the stuffed animal.
Explanation Generation	
Statement	*__Coke__ is made of **alcohol**.*
GPT2-FT	Coke is a drink.
GPT2-OMCS-FT	Coke is not a liquid.
GRF	Coke is made from **corn**.
Statement	*She **cut** up a **blanket**.*
GPT2-FT	A blanket is not sharp enough to cut.
GPT2-OMCS-FT	A blanket is too small to be cut.
GRF	Blankets are too **soft** to be cut.

Table 8: Case study on the test set of three datasets. Words in blue denote source concepts in the input contexts while words in orange are the associated concepts generated by the GRF.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Ex-*

trinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4220–4230. Association for Computational Linguistics.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. 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 4762–4779. Association for Computational Linguistics.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2019. Question answering by reasoning across documents with graph convolutional networks. 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 2306–2317. Association for Computational Linguistics.

Wenhu Chen, Wenhan Xiong, Xifeng Yan, and William Yang Wang. 2018. Variational knowledge graph reasoning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1823–1832. Association for Computational Linguistics.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, Luke Vilnis, Ishan Durugkar, Akshay Krishnamurthy, Alex Smola, and Andrew McCallum. 2018. Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Jian Guan, Fei Huang, Minlie Huang, Zhihao Zhao, and Xiaoyan Zhu. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Trans. Assoc. Comput. Linguistics*, 8:93–108.

Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. 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 6473–6480. AAAI Press.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

James N Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2016. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114 13:3521–3526.

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

pages 110–119. The Association for Computational Linguistics.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2829–2839, Hong Kong, China. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xi Victoria Lin, Richard Socher, and Caiming Xiong. 2018. Multi-hop knowledge graph reasoning with reward shaping. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3243–3253. Association for Computational Linguistics.

Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2019. Knowledge aware conversation generation with explainable reasoning over augmented graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1782–1792. Association for Computational Linguistics.

Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and Songlin Hu. 2019. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. *CoRR*, abs/1909.05311.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1506–1515. Association for Computational Linguistics.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. 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 845–854. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James F. Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 839–849. The Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Lin Qiu, Yunxuan Xiao, Yanru Qu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. 2019. Dynamically fused graph network for multi-hop reasoning. 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 6140–6150. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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.

Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 593–607. Springer.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083. Association for Computational Linguistics.

R. Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC*

2012, Istanbul, Turkey, May 23-25, 2012, pages 3679–3686. European Language Resources Association (ELRA).

Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. *CoRR*, abs/1806.02847.

Yi-Lin Tuan, Yun-Nung Chen, and Hung-yi Lee. 2019. Dykgchat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1855–1865. Association for Computational Linguistics.

Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha P. Talukdar. 2020. Composition-based multi-relational graph convolutional networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2019. Does it make sense? and why? A pilot study for sense making and explanation. 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 4020–4026. Association for Computational Linguistics.

Pengcheng Yang, Lei Li, Fuli Luo, Tianyu Liu, and Xu Sun. 2019. Enhancing topic-to-essay generation with external commonsense knowledge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2002–2012, Florence, Italy. Association for Computational Linguistics.

Zhi-Xiu Ye, Qian Chen, Wen Wang, and Zhen-Hua Ling. 2019. Align, mask and select: A simple method for incorporating commonsense knowledge into language representation models. *CoRR*, abs/1908.06725.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. 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.