Guiding Dialogue Agents to Complex Semantic Targets by Dynamically Completing Knowledge Graph

Yue Tan[1], **Bo Wang**[1]*, **Anqi Liu**[1], **Dongming Zhao**[2],
Kun Huang[2], **Ruifang He**[1], **Yuexian Hou**[1]
[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]AI Lab, China Mobile Communication Group Tianjin Co., Ltd.
{tanyue_098, bo_wang, anqi_liu }@tju.edu.cn

Abstract

In target-oriented dialogue, the representation and achievement of targets are two interrelated essential issues. In current approaches, the target is typically assumed to be a single object represented as a word, which makes it relatively easy to achieve through dialogue with the help of a knowledge graph (KG). However, when the target has complex semantics, the existing KG is often incomplete in tracking semantic relations. This paper studies target-oriented dialog where the target is a topic sentence. We combine the methods of knowledge retrieval and relationship prediction to construct a context-related dynamic KG, in which we can track the implicit semantic paths in the speaker's mind that may not exist in the existing KGs. In addition, we also designed a novel metric to evaluate the tracked path automatically. The experimental results show that our method can control the agent more logically and smoothly toward the complex target.

1 Introduction

Different from the open-domain and task-oriented dialog, the target-oriented dialog is a more challenging task that aims to achieve a global target through the dialog. This process cannot be decomposed into subtasks as in a task-oriented dialogue and is excepted to be semantically coherent and effective with fewer turns. Target-oriented dialog agents have a broad-based demand, e.g., psychotherapy (Sharma et al., 2020), conversational recommendation (Kang et al., 2019), and education (Clarizia et al., 2018), where the agent is expected to guide the dialog to a global target, e.g., a mental state, an item, and a knowledge point, respectively.

In general, the target of target-oriented dialog can be an entity (e.g., an item) or a topic (e.g., a knowledge point). The topic target is more challenging because of complex semantics, often simplified as keywords in existing works (Tang et al., 2019; Zhong et al., 2021a). In this way, existing approaches often require a knowledge Graph (KG) to retrieve relevant knowledge between the current dialog context and target keywords (Zhong et al., 2021a; Yang et al., 2022). Some latest work of target-oriented dialog also used the stored knowledge in LM to generate knowledge paths to assist dialog generation (Gupta et al., 2022).

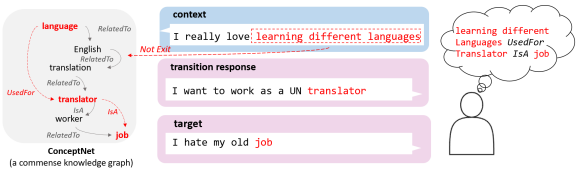

Figure 1: An example of using KG path in transferring topic to a target sentence. The key phrase "learn different languages" in the context is missed in the KG, and only "translator" appears in the human transition response. This happens because some relations in the human speaker's mind do not exist in the KG. If missed concepts and relations can be completed in kG, we can link the context and target with the transition response.

However, there are still issues for knowledge-based approaches to target-oriented dialogue: (1) The keywords are often ambiguous to represent complex target semantics. (2) KG knowledge is often insufficient. Concepts and relations for target-oriented processes in specific dialogue are often missed in widely used common KG (Zhong et al., 2021a; Yang et al., 2022), which results in failed or redundant long processes. For example, in Figure 1, the key phrase "learn different languages" can reflect contextual semantics better than a single concept. But it is not a node in the KG. In addition, the critical two-hop logic used by the speaker (language-translator-job) is missed in KG, while in the alternative long path, the concepts (e.g., "English", "worker") are redundant for the response generation. (3) KG path acquisition is challenging due to the large search space. Furthermore, com-

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2023, pages 6506–6518

plex target semantics requires more precise control over the space of knowledge selection, which is different from current works that use knowledge to enrich response generation without target restriction (Zou et al., 2021; Zhou et al., 2022) or only use the keyword as the target (Gupta et al., 2022).

To address these issues, in this work, we represent the target topic with a sentence instead of keywords. Subsequently, instead of using a static KG, we achieve the target sentence by reasoning on a dynamic KG. Before the response generation, the dynamic KG is generated based on static KG according to the dialogue context and the target sentence. This dynamic KG is expected to involve a more context-relevant and shorter path toward the target sentence. Specifically, besides the node and edges in the static KG, the additional dynamic nodes include key phrases in the dialog context. A relationship prediction model predicts the additional dynamic edges. To control the space of KG path selection more reasonably, in constructing the dynamic node, we use an extended "phrases bag" and a trained model respectively to ensure the diversity and relevance of nodes in the dynamic graph. In addition, we design an automatic metric for knowledge path evaluation, considering the convergence of path semantics with the context and target semantics.

Our main contributions are as follows:

(1) For guiding dialogues towards a given target sentence, we design a knowledge path generation method based on a dynamic KG. As far as we know, this is the first time relationship prediction has been used for multi-hop reasoning of topic transition in target-oriented dialogues.

(2) We propose an automatic metric to evaluate the quality of generated knowledge paths, considering the inference relationship between path fragment semantics and sentence semantics.

(3) We extracted a subset from the dialogue data set including hard cases where the target-oriented transition cannot be matched by a static KG path and verified the effectiveness of our method on it.

2 Related Work

Target-oriented dialogue agents In the study of target-oriented dialogue agents, a typical simplified task is keyword-guided dialog leading the dialog to a given keyword or a recommended item through multi-turn dialogue. The task is often divided into two stages (Tang et al., 2019; Qin et al., 2020; Zhong et al., 2021a), in which the first stage is to predict a next-turn keyword, and the second stage is keyword-based response retrieval. Instead of keywords, our work uses sentences with more complex semantics as the global target. In this direction, (Gupta et al., 2022) obtains SOTA performance using a pre-trained language model to generate multi-hop paths between a pair of concepts for transition response generation. Regarding data, (Sevegnani et al., 2021) propose a popular dataset for target-oriented dialog, which will be used in our work.

Commonsense Reasoning Recent approaches have realized the importance of commonsense reasoning in language generation, e.g., (Ji et al., 2020a) studied commonsense explanation generation. In this work, we follow the researches that utilize commonsense reasoning in generation models (Zhong et al., 2021b; Zou et al., 2021; Zhou et al., 2022). (Yang et al., 2022; Zou et al., 2021) select next-turn concepts from the static KG, conditioned on the dialogue context. Different from this kind of knowledge retrieval method, (Zhou et al., 2022; Gupta et al., 2022) generates implicit knowledge using a language model. (Becker et al., 2021) combines relation classification and target prediction for generating commonsense knowledge representations over text. Similarly to it, we also used the relation prediction method. Still, we use it to complement the knowledge graph to obtain multi-hop paths and combined knowledge retrieval to enhance controllability.

Commonsense Path Evaluation Most research that involves utilizing commonsense knowledge for tasks such as question answering (Kapanipathi et al., 2020) and commonsense reasoning (Lin et al., 2019) tend to use paths extracted from static knowledge graphs. However, the effectiveness of the knowledge paths is evaluated indirectly through the performance of downstream tasks in these work. (Becker et al., 2021) automatically evaluates the knowledge path through the similarity with the implicit knowledge in the dataset. Still, this method only works when annotated golden paths are provided in the dataset. We will address this issue with a novel automatic evaluation metric based on the semantic connection between dialogues and paths.

3 Methodology

3.1 Problem Statement

We frame the target-oriented response generation:

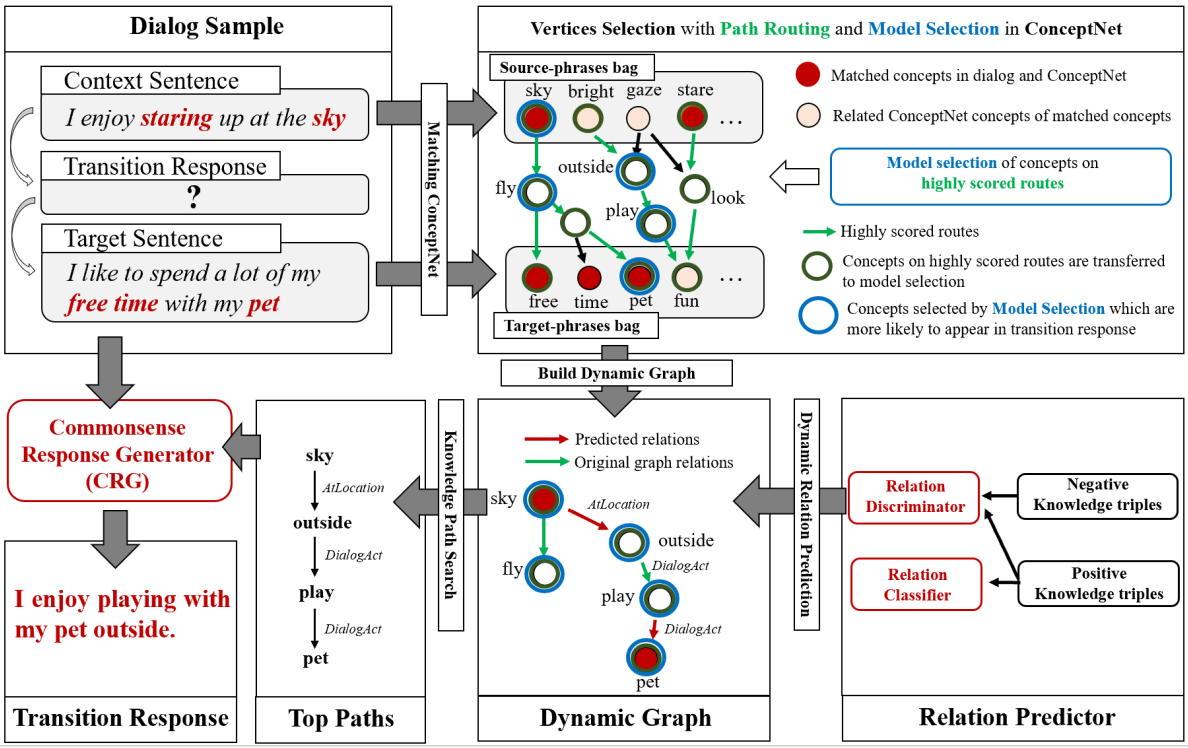

Figure 2: Framework of our model. A dynamic KG is completed for transition response generation through two steps: node selection and edge construction. The node selection includes path routing and model selection in ConceptNet, and the edge construction includes relation prediction and discrimination. Finally, the top connected multi-hop paths from the dynamic graph are sent to the commonsense response generator for response generation.

Given a dialog context c and a target t, a conditional language model learns to predict a transition response y. Our model finds a bridge path p on a dynamically acquired KG G. Then we use p, c and t to generate a proper y. The explanation of c, t, and y are as follows: **dialog context** c: A sentence that can represent the topic in the current dialog context. **target** t: A sentence representing the target topic of the current dialog. **transition response** y: A sentence that logically connects the semantics of c and t.

3.2 Method Overview

Figure 2 shows the overview architecture of our proposed model. Before using the pre-trained language model to generate a transition response, we built a dynamic graph to obtain the path, including two steps of node selection and edge construction. In the node selection, we ensure the diversity of nodes in the dynamic graph through extended "source-phrases bag" and "target-phrases bag" and ensure the contextual relevance of nodes through path routing and model selection. In the edge construction, we use a relation prediction model to complement the static graph. Finally, to generate transition responses, we generate multi-hop paths from the dynamic graph and send them into Commonsense Response Generator (CRG model) based on a pre-trained GPT2. In order to automatically and unbiasedly evaluate the advantages of our paths, we design an automatic evaluation metric referring to the idea of the NLI task. First, each candidate path is divided into fragments. We suppose that for a path reasonably connected with the dialog, the source sentence should entail the start fragment of the path, and the target sentence should entail the end fragment of the path. With this idea, we construct positive and negative samples to train a classifier model for path evaluation.

3.3 Dynamic Graph Building

To both make full use of the existing knowledge in the KG and infer additional knowledge related to the current dialog, we combine knowledge retrieval and relation prediction to build a dynamic graph.

3.3.1 Dynamic Node Selection

This step ensures that the nodes in the dynamic graph are diverse and context related. We referred to the idea of using path routing and concept selection to deactivate nodes in (Ji et al., 2020a), and

made changes suitable for our tasks regarding path acquisition, path representation, concept word representation, etc.

Path Routing To obtain the initial candidate nodes, we heuristically retrieve multi-hop paths from the ConceptNet based on the context and target sentence. To include diverse candidate words, we use an extended "source-phrases bag" as the start of the path, which contains both key phrases in the source sentence and the most semantically similar neighbor phrases. Similarly, the "target-phrases bag" is the end of the path. Then the path routing propagates the scores along the paths to each candidate concept. For each retrieved path p, we calculate a score $s(p)$ according to a soft-matching procedure. Each p is converted into a natural language form, and then we use Sentence-BERT (Reimers and Gurevych, 2019) to measure $s(p)$ as the p's semantic similarity with the dialogue sentences. Finally, we get the routing score of a candidate concept c by the average $s(p)$ of all the paths passing c (i.e., $\mathbf{P_{v_1 \to c \to v_2}}$).

$$s(c) = \frac{1}{|\mathbf{P_{v_1 \to c \to v_2}}|} \sum_{\mathbf{p} \in \mathbf{P_{v_1 \to c \to v_2}}} s(\mathbf{p}) \quad (1)$$

A high routing score of a c indicates that the paths through c are highly related to the dialogue, so the concept word is important for this context. Finally, we preserve $\mathcal{V}_{s \to t}$ with top-$K1$ routing scores.

Model Selection For all concepts in $V_{s \to t}$, we use the sentence representation to query each concept representation by taking the dot-product attention and calculating the selection probability with supervision from concepts in gold response $\mathcal{C}_{s \to t}$:

$$P(c|s) = \sigma(h_c W h_x^T) \quad (2)$$

$$\begin{aligned} \mathcal{L}_{\text{concept}} = & -\sum_{\mathbf{c} \in \mathcal{V}_{\mathbf{x} \to \mathbf{t}}} \mathbb{I}\left(\mathbf{c} \in \mathcal{C}_{\mathbf{s} \to \mathbf{t}}\right) \log P(c \mid \mathbf{x}) \\ & + \left[1 - \mathbb{I}\left(\mathbf{c} \in \mathcal{C}_{\mathbf{s} \to \mathbf{t}}\right)\right] \log[1 - P(c \mid \mathbf{x})] \end{aligned} \quad (3)$$

where h_c is the concept representation encoded by GloVe, h_x is the concatenated representations of the source and target sentence encoded by GRU. W is a trainable parameter matrix. $\mathbb{I}\left(\mathrm{c} \in C_{\mathrm{s} \to \mathrm{t}}\right)$is an indicator function taking the value 1 iff $\mathbf{c} \in \mathcal{C}_{\mathbf{s} \to \mathbf{t}}$ and 0 otherwise. Finally, the bridge concepts with top-$K2$ $P(c|x)$ and the sentence pair's key phrase serve as the dynamic graph's nodes.

3.3.2 Dynamic Edge Construction

Our dynamic graph first inherits existing edges in KG and then uses a relation prediction model and relation discriminator to construct dynamic edges.

Relation Prediction and Discrimination We trained a relation prediction model to add edges to the dynamic graph. Given any pair of unconnected concepts, the model predicts and judges whether they can be connected. Specifically, we fine-tune a pre-trained language model DistilBERT on gold knowledge triples by masking the relations and treating it as a multi-classification task. To adapt to our tasks and minimize the limitation of incomplete knowledge, we filter and expand the training data (detailed in Section 4.1). Using the same training data, we also train a relation discriminator to ensure the predicated edges further.

3.4 Knowledge Path Search

Subsequently, we connect a pair of phrases from the source and target sentence using multi-hop paths. Specifically, assuming the source and target consist of m and n key phrases, we take any of the $m * n$ pairs of key phrases as the start and the destination to find paths within three hops in the dynamic graph obtained in 3.3. Finally, we use the top paths with low perplexity and high diversity scores for the transition response generation. This way, selected paths contain less irrelevant and redundant information while ensuring diversity and logicality.

3.5 Training the CRG model

Inspired by (Gupta et al., 2022), we send the final path in 3.4 to the Commonsense Response Generator (CRG) model together with the sentence pair to generate a transition response. The CRG model (GPT-2 based) is trained as a conditional model with the following input sequence: "[context]$source\ sentence$[target] $target\ sentence$ [knowledge] $knowledge\ path$ [response] $transition\ response$ ". We train the CRG model by minimizing the log-likelihood loss of the transition response.

3.6 Novel Evaluation of Transition Path

A good transition path should take into account the semantics of both the source and the target sentence and contains as little redundant information as possible. However, there is no annotated golden path in the corpus, and multiple reasonable paths may exist. We propose an automatic metric without

Dialog	Source: Response: Target:	I do not like to **cook**. I actually love to cook, but sharing the **kitchen** with three roommates makes it difficult. I want to get my **own place**.
Positive Path Fragment	Source→Response Response→Target	cook *uses* kitchen kitchen *is a* place
Negative Path Fragment	Source→Random Random→Target	cook *motivated by goal* create landmark *at location* a place

Table 1: Cases for training PATH-COHERENCE metric. We use the sentence-path fragment pair extracted from the corpus as positive samples and construct the same number of negative samples

golden references. Our primary hypothesis is that the semantics of the context and target sentence should entail the information of the start and the end fragment of the path, respectively. The proposed metric PATH-COHERENCE is based on a classification model trained to classify a sentence-path fragment pair are logically coherent or irrelevant. Formally, for a sentence s, a path fragment p_f, letting $\text{conf}_{\text{class}}\ (s, p_f)$ represent the model's probability mass assigned to the predicted NLI class after softmax(This is similar to the UNLI concept proposed in (Chen et al., 2020), i.e. we do not directly use classification labels), the function is defined as $NLI_{score}\ (s, p_f) =$

$$\begin{cases} 1 * \text{conf}_{\text{entailment}}\ (s, p_f) & \text{if coherent} \\ 0 & \text{if irrelevant} \end{cases} \quad (4)$$

For a complete path, we define the first triplet of the path as its start fragment p_{f_s} and the last triplet as its end fragment p_{f_t}. Then the PATH-COHERENCE of a path can be calculated as $NLI_{score}\left(s_s, p_{f\_s}\right) + NLI_{\text{score}}\ \left(s_t, p_{f\_t}\right)$,where s_s and s_t represent the source sentence and the target sentence respectively.

We use the transition paths from the golden responses to create positive samples for training. We identify its knowledge path through a hard-matching process with context c, target t, and response y (Table 1). Specifically, this process first identifies the key phrases in the sentence. If the key phrases of two adjacent sentences are directly connected in ConceptNet, the sentence and path fragment pair is regarded as a positive training sample. For the negative sample, we use the concepts in the "phrases bag" (mentioned in 3.2.1) of the sentence as the head or tail to randomly select the triples with different relationships in KG from the positive sample. The negative sample constructed in this way has a weak correlation with the dialogue, so it can better guarantee the model's discrimination.

4 Experiment

4.1 Dataset

For the relation predictor training, we use the CN-100k benchmark dataset (Li et al., 2016), based on the OMCS subpart of ConceptNet. The dataset comprises 37 relation types, 100k relation triples in the train set, and 1200 triples in the development and the test set, respectively. We extract a subset including 15 relationships that are most suitable for topic transition (detailed in the appendix). Intuitively, the knowledge triplets implied in the dialogue corpus that does not exist in the relation prediction training data, especially those with high frequency, actually reflect the commonsense logic of people in the real dialogue. With this idea, we filtered the concept pairs whose frequency of occurrence in two adjacent sentences is higher than a threshold in the OTTers corpus and defined their relationship as "DialogAct" to form new knowledge triplets. Finally, the dataset covers 102178 triples for training, 1236 triples for development, and 1245 triples for testing.

We use two datasets to test the transition response generation: 1) Otters (Sevegnani et al., 2021) contains instances with context-target-transition response triplets. It consists of two sets of splits. The Out-Of-Domain (OOD) split ensures that none of the context-target pairs in the test set are present in the train set. In the In-Domain (ID) split, one of either the context or the target in each pair in the test set is allowed to appear in the train set. 2) Augmentation-DailyDialog is similar to OTTers, which is constructed by (Gupta et al., 2022) from DailyDialog (Li et al., 2017). This data is noisier because of too many turns, sentence fragmentation, and serious overlap between transition response and target sentences.

To build a more challenging task, we also extracted a sub-dataset from OTTers, called "Discrete-OTTers"[1], which contains difficult cases for topic transition where the three golden transition responses corresponding to the dialog cannot match a connected path in ConceptNet.

[1]This dataset and code will be published in https://github.com/tanyue2019/ACL-Pro

4.2 Baselines

We compare our model with three groups of baselines: General generating model without additional knowledge (GPT-2), concept-guided models (Concept-Predict, MultiGen), and path-guided models (Static, CODA, TBS-Path). Implementation details of baselines are in Appendix A.

GPT-2 (Radford et al., 2019), a pre-trained GPT–small language model fine-tuned on Otters data. Conditions on the context and target sentences to generate the transition response.

Concept-Predict leverages concept prediction strategy in(Zhong et al., 2021a). The predicted concepts are filtered based on closeness to the target.

MultiGen (Ji et al., 2020b) combines the vocabulary distribution generated by the underlying GPT-2 model with a concept distribution from a commonsense knowledge base (ConceptNet).

Static uses ConceptNet to extract paths between concepts from sentence pairs and generate a response using a generation model.

CODA (Gupta et al., 2022) proposes a method to generate multi-hop bridging paths for target-oriented response generation.

TBS-Path first externalizes implicit commonsense knowledge based on the dialog context like Zhou et al. (2022) and uses the knowledge to generate responses.

Ablation models:

StaticRelation variant that uses the multi-hop connection in ConceptNet to replace the edge predicted by the relationship prediction model in test paths. If no connections are within 4 hops, use "[SEP]" to connect.

RandomConcept variant that randomly selects top-$K2$ neighbor nodes within two hops in the knowledge map of context concepts to construct the dynamic graph.

FewerHops variant that uses a shorter path for transition response generation.

4.3 Evaluation Metrics

4.3.1 Paths Evaluations

Automatic Evaluation Perplexity (PPL) measures the smoothness of the path, and our designed PATH-COHERENCE (Section 3.6) measures the correlation and coherence between the path and sentence.

Human Evaluation For randomly selected 100 generated paths and their corresponding sentence pair, we ask annotators to judge 1) Relevance: Is this path relevant and coherent to the context of the sentence pair? 2) Makes sense: Does the path makes sense? Four annotators with an NLP background score the paths in 1, 2, 3, higher is better.

Source Topic:	I enjoy staring up at the **sky**
Response:	I love watching the sky while **walking my dog**.
Target Topic:	I like to spend a lot of my **free time** with my **pet**.
Manual Path:	**sky** -*LocatedAt*-> outside -*RelatedTo*-> nature <-*RelatedTo*- animals <-*IsA*- **dog** -*IsA*-> **pet**
Source Topic:	i really love **learning** different **languages** and have been **studying** them for years.
Response:	I want to work as a UN **translator**.
Target Topic:	i **hate** my old **job**.
Manual Path:	**language** -*RelatedTo*-> English -*RelatedTo*-> translation -*RelatedTo*-> **translator** -*IsA*-> worker -*RelatedTo*-> **job**
Source Topic:	i tell **jokes** on stage.
Response:	Being a **comedian** has opened up a lot of **dating** opportunities for me.
Target Topic:	i **date** a lot of **girls**.
Manual Path:	**jokes** <-*RelatedTo*- **comedian** -*RelatedTo*-> comedy <-*HasPrerequisite*- entertaining someone <-*UsedFor*- going to a film -*UsedFor* -> **dating** -*RelatedTo*-> **date**

Table 2: Examples of Discontinuous Paths in the Knowledge Graph Reflected in Dialogue Logic

Model	PPL	PC	Relevance	Makes Sense
Static	8.79	16.51	1.06	1.10
TBS-Path(Zhou et al., 2022)	7.44	28.72	1.56	1.45
CODA(Gupta et al., 2022)	9.15	29.91	1.78	1.69
Ours	7.59	**40.87**	**2.28**	**2.15**
kappa (The agreement among the annotators.)			0.51	0.55

Table 3: Evaluation for path quality. Our path has significant advantages in PC results. The consistency between PC metric and human evaluation also proves the rationality of this metric design.

4.3.2 Response Evaluations

Automatic Evaluation We report standard automated metrics such as BLEU(Papineni et al., 2002)[2] , METEOR(Banerjee and Lavie, 2005),ROUGE-L(Lin, 2004) and BertScore(Zhang et al., 2019). Word-overlap metrics do not correlate well with human judgements(Liu et al., 2016). So we also adopted the metric TARGET COHERENCE designed by(Gupta et al., 2022), which does not require human references but evaluates the coherence of replies based on a trained classification model.

Human Evaluation Annotators are requested to evaluate the transition response on the follow-

[2]SacreBLEU (Post, 2018) provides hassle-free computation of shareable, comparable, and reproducible BLEU scores. The calculation is carried out using multiple references from the dataset

	OTTer-ID					OTTer-OOD				
	BLEU	METEOR	ROUGE-L	BS-f1	TC	BLEU	METEOR	ROUGE-L	BS-f1	TC
GPT2(Radford et al., 2019)	10.44	16.93	17.79	76.91	41.39	10.06	17.71	19.06	77.65	41.79
Concept-Predict(Zhong et al., 2021a)	14.91	15.89	19.60	77.67	35.46	12.89	15.69	19.80	78.13	38.77
MultiGen(Ji et al., 2020b)	18.45	17.46	19.82	78.15	47.87	13.94	17.73	20.91	78.02	45.89
Static	11.93	18.13	17.49	76.82	41.27	13.19	19.87	20.08	78.02	48.70
CODA(Gupta et al., 2022)	16.05	16.61	19.83	77.60	46.84	14.76	16.64	20.76	77.97	49.82
TBS-Path(Zhou et al., 2022)	13.98	17.95	19.31	78.01	49.05	14.63	18.02	20.53	78.41	46.67
Ours	**20.14***	18.11	**21.12***	78.01	**52.98***	**18.08**	18.78	22.18	78.67	**51.44***
Ours-StaticRelation	18.45	18.47	20.35	78.06	49.41	15.83	18.49	21.76	78.41	48.72
Ours-RandomConcept	19.49	18.08	20.41	77.61	49.23	17.61	18.80	21.78	78.52	50.41
Ours-2hop	19.17	18.04	20.75	77.91	49.25	18.05	19.18	22.53	78.79	50.87

Table 4: Automatic evaluation on OTTers. We also present results for our model's ablations. The results of our model on most reference-based metrics and model-based metrics exceed the baselines. (t-test with p-value < 0.05)

	BLEU	METEOR	ROUGE-L	BS-f1	TC
GPT2(Radford et al., 2019)	8.20	19.78	21.74	75.06	74.37
Concept-Predict(Zhong et al., 2021a)	6.09	17.69	18.19	74.85	71.83
MultiGen(Ji et al., 2020b)	2.83	14.75	14.60	73.13	76.53
Static	9.87	21.09	21.89	74.99	74.74
CODA(Gupta et al., 2022)	8.24	19.09	19.53	74.08	75.84
TBS-Path(Zhou et al., 2022)	9.92	21.78	21.93	74.73	77.21
Ours	**12.61***	**23.84***	**24.49***	**75.58***	**77.46***

Table 5: Automatic evaluation on Augmentation-DailyDialog

	BLEU	METEOR	ROUGE-L	BS-f1	TC
Static	9.93	20.49	17.68	75.34	43.90
CODA(Gupta et al., 2022)	11.45	16.77	19.67	76.33	43.08
TBS-Path(Zhou et al., 2022)	10.49	16.54	18.82	73.09	45.15
Ours	**14.08***	**20.03***	**20.10***	**77.94***	**54.11***

Table 6: Automatic evaluation of path-based methods on Discrete-OTTers. The performance of our model on this difficult data subset is still significantly better than that of other path-based models, which shows the effectiveness of our path-acquisition method.

Model	Coherent	Sensible	Informative
Ours vs. GPT-2	64%	60%	59%
Ours vs. Static	55%	57%	49%
Ours vs. MultiGen	56%	55%	59%
Ours vs. Concept-Predict	65%	63%	62%
Ours vs. CODA	57%	60%	50%
Ours vs. TBS-Path	62%	61%	56%

Table 7: Human evaluation through pairwise comparison between our model and baselines (using sign test, p<0.05). Our model is preferred in coherence and sensible criteria while being comparably informative.

ing criteria: (1) Smooth: rate whether the response serves as a smooth transition between the dialogue context and target. (2) Sensible: whether the transition response makes sense in itself, i.e., it is grammatical and logically coherent. (3) Informative: how much informative content a transition response carries. Four annotators with an NLP background compare transition responses from two models.

4.4 Preliminary Experiment

The preliminary experiment examines the model's ability to use discontinuous paths in the KG fully. We extracted such cases from the dataset: the key phrases in their source sentences, transition response sentences, and target sentences are not directly connected in the KG. We manually check the KG to annotate a reasonable transfer path for these cases. As shown in Table 2, the logical connection in dialogue is probably just a few discontinuous hops in the long path of the graph. If additional edges connect these discontinuous nodes, the transition path will be more efficient.

4.5 Results

4.5.1 Paths Evaluations

In Table 3, the PPL results indicate that our paths have good fluency, which means they can be better accepted by the language model to generate transition responses. The significant advantage of the PC metric proves that our method is effective in obtaining context-related paths. The results of the manual evaluation are similar to those of PC, which shows that the automatic evaluation metric we designed is largely consistent with human judgments.

4.5.2 Response Evaluations

Automatic Evaluation As shown in Tables 4 and 5. on two datasets, the results of our model on most reference-based metrics and model-based metrics exceed the baselines. This indicates the advantage that the path we input to the model is semantically connected with less context-independent information. In Table 6, we provide the evaluation results of our model and three path-based baselines on the discrete OTTers we extracted. As mentioned earlier, this dataset is challenging because it is dif-

Source: i work in a library; **Target:** i had cows as pets growing up	
Static	**Response:**My cat jumps on my good book. **Path:** library *causes* read *related to* eyes *is a part of* cat *is used for* pet
CODA	**Response:** I grew up in a library. textbfPath:library *is the location which has* books *not capable of* grow.
TBS-Path	**Response:**I love to work for a dog. **Path:** work *has prequisite* not work *is a subevent of* have dog *motivated by goal* grow
Ours	**Response:** My school *is located in* a rural area. **Path:** library *is at location* school *dialog act* area *is used for* pets growing

Source: my job helpe me teach kids; **Target:** education is a passion of mine.	
Static	**Response:** I teach kids for patience **Path:** job *related to* patience *related to* patient *related to* passion
CODA	**Response:**I want to have some sweet. **Path:**kid *desires* candy *is a dependency of* tasting sweet *causes* passion
TBS-Path	**Response:**I love teaching **Path:** teach kids *has a context* education
Ours	**Response:** I teach children how to learn. **Path:** teach kids *is used for* children *dialog act* learn *is used for* education.

Table 8: Case study on OTTers. Our method generates better transition responses with better knowledge paths, which have stronger contextual relevance and logical rationality.

ficult to transfer the topic of the dialogs in it. The results show that our method has good robustness for such difficult situations. This is because we can use the knowledge outside the knowledge graph to connect two sentences with far semantics and ensure contextual relevance.

Human Evaluation We collect 100 randomly selected data points from the test outputs on OTTers. The score in Table 7 is the percentage of times that our model is chosen as the better in pairwise comparison with its competitor. The results demonstrate that our outputs are preferred over the baselines, especially on "Smooth" and "Sensible".

4.5.3 Ablation Studies

Ablation results are shown in Table 4.

Can relation prediction effectively use discontinuous paths in the KG?

As shown in Table 4 that after replacing the relation in our test paths with the relation or multi-hop path in the KG, all metrics decrease significantly. We analyze the results and find that the length of the replaced path has increased by seven hops on average, and the path contains a lot of ambiguous relations, such as "related to". This verifies that we can efficiently connect nodes in the dynamic graph through relation prediction.

Can concepts filtering reduce redundant information in the path? After using the random top-$K2$ concepts within two hops as the nodes in the dynamic graph, the results are reduced, but the decline is not significant. We analyzed the test paths obtained by this method. We found that the relation prediction and discriminator in the model largely ensured that the final test path contained less redundant information. Specifically, due to random selection, most of the nodes in the dynamic graph are uncorrelated, so our relation discriminator mostly negates the results of relation prediction at this time. These irrelevant nodes are not connected in the dynamic graph.

How much path information do we need? We explore whether 3-hop paths provide more redundant information than 2-hop paths. In Table 3 (Ours-2hop), we can see little difference between the word-overlap metrics using the two-hop path and the three-hop path. Still, the TC result has decreased, which proves that it is difficult to achieve a smooth transition by relying on an intermediate word. Therefore, we finally used the 3-hop paths as the test data.

4.5.4 Case Study

We compare our method with the other three path-based methods(Table 8). It can be seen from two examples that the path obtained by the Static contains many fuzzy relations and irrelevant concepts. Thanks to the training on the path data related to the response, the path obtained by CODA is better than the former. However, it still exists in the information irrelevant to the dialogue context. The path obtained from the TBS-Path contains more information duplicated with the conversation statement. The above problems lead to poor responses to these models, and there are some unreasonable points in the topic transfer logic. The path semantics obtained by our model is clear and logical, resulting in better responses. It is noted that the "DialogAct" relation we added also played a good role.

5 Conclusion

For effectively guiding the dialog to a target sentence, we propose to make full use of discontinuous or even non-existent paths in the knowledge graph. We combine knowledge retrieval and relationship

prediction by building a dynamic KG, which helps to obtain a path closer to the implicit logic in the speaker's utterances when transferring topics to the target sentence. In addition, we also designed an automatic metric to evaluate the quality of the knowledge path for semantic transfer. Both automatic and human evaluation verify the superiority of the proposed method in searching knowledge paths and subsequently generating transition responses compared with SOTA baselines, benefiting from the better logic and contextual relevance of the paths from the dynamic graph. In the future, we will explore the application of our method to multi-turn target-oriented dialogue.

Limitations

The current dialogue system still has some limitations. For example, although the current CRG model can make the output contain the key concept words in the knowledge path, due to the large scale of the pre-training model, the output semantics of the current method are still not very interpretable and controllable. A feasible way is to explore new fine-tuning methods to approach high-level semantic style control.

In addition, our current dialogue system lacks human qualities such as empathy, factual correctness judgment, and moral common sense representation. A key breakthrough is to explore a goal-oriented dialogue dataset with richer dimensions.

Ethics Statement

A target-oriented dialogue system may have the risk of misusing it to guide users to malicious topics actively. Since the proposed system tries to ensure relevance with the dialogue context in the application deployment, the possibility of the above misuse is small on the premise of checking the training corpus.

All models in this paper are trained on the public corpus. The used datasets do not contain personal information or unethical language. We also ensure the anonymization of the human evaluation.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62272340, 61876128, 61876129, 62276187, 61976154, 61402323), State Key Laboratory of Communication Content Cognition (Grant No.A32003).

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Maria Becker, Katharina Korfhage, Debjit Paul, and Anette Frank. 2021. Co-nnect: A framework for revealing commonsense knowledge paths as explicitations of implicit knowledge in texts. *arXiv preprint arXiv:2105.03157*.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online. Association for Computational Linguistics.

Fabio Clarizia, Francesco Colace, Marco Lombardi, Francesco Pascale, and Domenico Santaniello. 2018. Chatbot: An education support system for student. In *International Symposium on Cyberspace Safety and Security*, pages 291–302. Springer.

Prakhar Gupta, Harsh Jhamtani, and Jeffrey P Bigham. 2022. Target-guided dialogue response generation using commonsense and data augmentation. *arXiv preprint arXiv:2205.09314*.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, and Minlie Huang. 2020a. Generating commonsense explanation by extracting bridge concepts from reasoning paths. *arXiv preprint arXiv:2009.11753*.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. 2020b. Language generation with multi-hop reasoning on commonsense knowledge graph. *arXiv preprint arXiv:2009.11692*.

Dongyeop Kang, Anusha Balakrishnan, Pararth Shah, Paul Crook, Y-Lan Boureau, and Jason Weston. 2019. Recommendation as a communication game: Self-supervised bot-play for goal-oriented dialogue. *arXiv preprint arXiv:1909.03922*.

Pavan Kapanipathi, Veronika Thost, Siva Sankalp Patel, Spencer Whitehead, Ibrahim Abdelaziz, Avinash Balakrishnan, Maria Chang, Kshitij Fadnis, Chulaka Gunasekara, Bassem Makni, et al. 2020. Infusing knowledge into the textual entailment task using graph convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8074–8081.

Xiang Li, Aynaz Taheri, Lifu Tu, and Kevin Gimpel. 2016. Commonsense knowledge base completion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1445–1455.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually

labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. *arXiv preprint arXiv:1909.02151*.

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

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

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*, pages 311–318.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Jinghui Qin, Zheng Ye, Jianheng Tang, and Xiaodan Liang. 2020. Dynamic knowledge routing network for target-guided open-domain conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34 Issue 05, pages 8657–8664.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Karin Sevegnani, David M Howcroft, Ioannis Konstas, and Verena Rieser. 2021. Otters: One-turn topic transitions for open-domain dialogue. *arXiv preprint arXiv:2105.13710*.

Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*.

Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5624–5634.

Zhitong Yang, Bo Wang, Jinfeng Zhou, Yue Tan, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2022. Topkg: Target-oriented dialog via global planning on knowledge graph. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 745–755.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Peixiang Zhong, Yong Liu, Hao Wang, and Chunyan Miao. 2021a. Keyword-guided neural conversational model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35 Issue 16, pages 14568–14576.

Peixiang Zhong, Di Wang, Pengfei Li, Chen Zhang, Hao Wang, and Chunyan Miao. 2021b. Care: Commonsense-aware emotional response generation with latent concepts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14577–14585.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2022. Think before you speak: Explicitly generating implicit commonsense knowledge for response generation. In *ACL 2022*.

Yicheng Zou, Zhihua Liu, Xingwu Hu, and Qi Zhang. 2021. Thinking clearly, talking fast: Concept-guided non-autoregressive generation for open-domain dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

A Appendix

A.1 Implementation Details

Model training: The pre-trained models we used are from Huggingface library[3]. To construct the dynamic graph process, we set K1=100, K2=20, and use GloVe embedding of size 300 (Pennington et al., 2014) during node selection. When training the relation prediction model and the relation discrimination model, we finetune DistilBERT for ten epochs with batch size=64, learning rate=$1 * 10^{-5}$, and accumulate grad batches=4.

The CRG model is based on GPT-2 small architecture. We use a batch size of 16 and accumulate grad batches=2 for GPT-2 models. We use AdamW optimizer with an initial learning rate of $2 * 10^{-5}$.

Finally, our PATH-COHERENCE model is also based on DistilBERT. We set the batch size=64 and use AdamW optimizer with an initial learning rate of $2 * 10^{-5}$. The accuracy of our classification model for this metric has reached over 90%

[3]https://huggingface.co/

Relation prediction dataset: When inheriting the edges in the static knowledge graph and filtering the training data of the relation prediction model, we removed some very unusual relationships, merged the relationships with similar semantics, and finally retained AtLocation, CapableOf, Causes, MotivatedByGoal, Desires, HasProperty, HasSubevent, HasPrerequisite, IsA, MadeOf, NotCapableOf, PartOf, UsedFor, ReceivesAction, HasA.

Discrete-OTTers dataset: We use the key phrase in the source sentence as the start and the key phrase in the target sentence as the end to find two hop paths in the static ConceptNet. If all paths do not include the key phrase in the bridge sentence in the corpus, we consider this conversation to be a separate case.

A.2 Training Details of Baselines

Training Concept-Predict leverages concept prediction strategy in(Zhong et al., 2021a). Following (Gupta et al., 2022) The input to the model is the context and target, and it predicts a single concept based on closeness to the target. The concept is then fed as input to a CRG model along with the context and target sentences.

Training Static It is a commonly used method to obtain paths from a fixed knowledge graph. Specifically, for a sentence pair, we start with the keywords in the source sentence and end with the keywords in the target sentence to find paths in the ConceptNet. To ensure that all test cases can find paths in this way, we set the maximum path length to be no more than 4. Finally, we also filter the path into a final Commonsense Response Generator based on PPL and diversity.

Training TBS-Path leverages the idea of generating implicit knowledge based on the context in (Zhou et al., 2022). Specifically, we use the path training data provided by(Gupta et al., 2022) because they all adopt the method of directly generating the path and use the pre-trained GPT2 to train the path generation model. The input to the model is the combination of source sentence and target sentence and the output of the model is the corresponding path. Finally, like our model, the path is sent to a Commonsense Response Generator for reply generation.

For MultiGen and CODA, we adopted the training methods provided in Sevegnani et al. (2021) and Gupta et al. (2022) respectively.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 6 Limitations

☑ A2. Did you discuss any potential risks of your work?
Section 7 Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1 Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☒ B1. Did you cite the creators of artifacts you used?
Left blank.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Left blank.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

☒ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☑ Did you run computational experiments?

Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A.1

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A.1

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
We mainly analyze and compare the performance metrics of the model. Further in-depth statistical analysis of the results will be conducted in future work.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix A.1

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 4.3

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
We trained the announcers in the form of meetings and displayed the main descriptions in section 4

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
section 4.3

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
The results of human evaluation are presented in tabular form

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No data collection

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
section 4.3