GraphDiffs: Graph Modeling with Differential Sequence for Document-Grounded Conversation

Anonymous ACL submission

Abstract

Knowledge grounded dialogue systems need to incorporate natural transitions between knowledge for dialogue to flow smoothly. Current systems not only lack good structured representations for knowledge that span multiple documents, but also effective algorithms that utilize such resources. We design a Co-Referential Multi-Document Graph(CoRM-DoG) that seamlessly captures inter-document correlations and intra-document co-referential knowledge relations. To best linearise this static graph into sequential dialogues, we contribute a Graph Modeling with Differential Sequence (GraphDiffs) method for knowledge transitions in dialogue. GraphDiffs performs knowledge selection by natively accounting for contextual graph structure and introducing differential sequence learning to effectively learn multi-turn knowledge transitions. Our analysis shows that GraphDiffs based on CoRM-DoG significantly outperforms the current state-of-the-art by $\mathbf{9.5}\%$ and $\mathbf{7.4}\%$ on two public benchmarks, WoW and Holle-E, where the modeling of co-reference and differential sequence are critical factors for its success.

1 Introduction

Document grounded conversations (Moghe et al., 2018; Dinan et al., 2018; Feng et al., 2021), as one type of knowledge grounded dialogue, leverage natural text-based knowledge sentences from documents to generate informative dialogue responses. The dialogues flows in these conversations reflect the logical relations in documents and also link topics of these documents through their commonsense relations. Document grounded dialogue system is usually divided into two sub-tasks (Dinan et al., 2018), knowledge selection and response generation given the dialogue history. Knowledge selection, also known as knowledge transition (Meng et al., 2020), is a crucial sub-task because it is equivalent to the dialogue flow management and also determines the content of the generated responses (Moghe et al., 2018; Dinan et al., 2018).

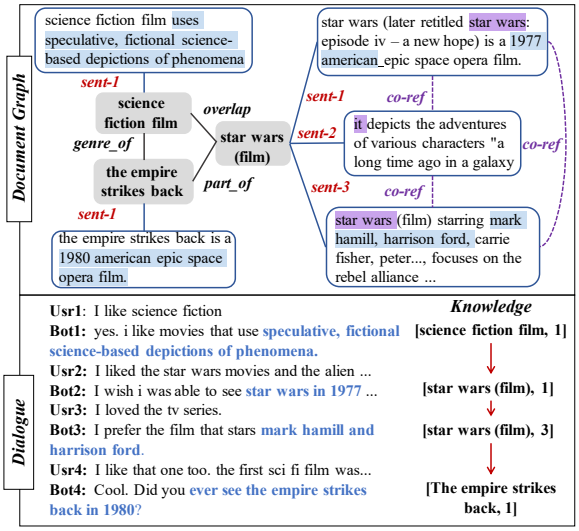


Figure 1: Co-Referential Multi-Document Graph (CoRM-DoG). Grey rounded rectangles are topic nodes of documents, interconnected by commonsense and word overlap relations. Rounded rectangles with blue frames are knowledge sentences in documents, connecting with the topic node by edge sentence order, denoted by *sent 1/sent 2* and so on. Knowledge sentences are inter-connected by co-reference relation(purple dash lines). Lower part are conversation and the knowledge sequence (bold word in '[]', numbers are the order of this knowledge in its topic) used in each turn.

Most of the existing research (Lian et al., 2019; Zheng et al., 2020; Zhao et al., 2020) for document-grounded conversations treats knowledge candidates independently, neglecting the critical relationships between knowledge candidates when performing knowledge selections. A recent work (Wu et al., 2021), proposes to encode knowledge candidates at the passage-level so that local relationships of knowledge are implicitly incorporated. However, the relations between documents and the more complex structure information contained within each document are continued to be ignored. In fact, as illustrated by the example in Figure 1, one single conversation encompasses several relevant topics

and knowledge sentences, with each knowledge sentence pertaining to a certain topic. The relationships between these topics typically dictate when the dialogue topic is changed, whereas the correlations between knowledge matter when the topic is maintained in multiple turns, which necessitates the exploration of the structural representations of knowledge spanning multiple relevant documents.

Although incorporating structure-aware representations of knowledge provide hints and inherent logic for knowledge transition, it is still non-trivial to adapt them into the multi-turn sequential dialogue flow management due to the lack of dialogue sequence information. As shown on the right bottom of Figure 1, the knowledge used by dialogue appears in a certain order. In general, historical knowledge sequences in different orders imply distinct subsequent knowledge and historical knowledge with varying distances to the current turn usually contributes differently to current knowledge selection. As a result, capturing the sequential pattern of knowledge transition is a critical complement to the structure-based knowledge representation and could further assist to pinpoint the right knowledge.

Based on the above concerns, we build the Co-Referential Multi-Document Graph (CoRM-DoG) to gather the structured information contained in documents. The co-reference knowledge correlations in documents are presented, as well as topic relations across multiple documents. Besides, we develop a novel Graph Modeling with Differential Sequence (GraphDiffs) method for learning the topic and knowledge transitions in document grounded conversations. Specifically, we employ a residual relational graph neural network to fully comprehend the document graph structure, which is later enhanced by a novel differential sequence learning method to perform knowledge selection, emphasizing the inter-turn knowledge shift sequence in dialogue history.

To sum up, our contributions in the paper can be summarized as follows. (1) We are the first to utilize the document graph structures for document grounded conversations and show that our CoRM-DoG performs the best by empirically comparing with other graph construction methods. (2) To fully adapt the CoRM-DoG into document grounded conversation flow management, we propose a novel GraphDiffs method to seamlessly incorporate the graph structured information built from documents and the differential knowledge sequence in a dialogue history. (3) We demonstrate our GraphDiffs based on CoRM-DoG outperforms the SOTA with significant margins over $\mathbf{9.5}\%$ and $\mathbf{7.4}\%$ on knowledge selection accuracy, for two public datasets, WoW and Holl-E, respectively.

2 Related Work

Document grounded Conversation. Initially, document grounded dialogue systems (Ghazvininejad et al., 2018) generate responses by copying words from the external documents. With the introduction of document grounded conversations where knowledge for each dialogue is annotated (Dinan et al., 2018; Moghe et al., 2018), the tasks of knowledge selection and response generation can be separated. Most of the previous works on document grounded conversation optimized knowledge selection by matching the dialogue context with each knowledge separately (Dinan et al., 2018; Lian et al., 2019; Zheng et al., 2020; Zhao et al., 2020). A recent work (Wu et al., 2021) proposed to treat knowledge at the passage level so that knowledge relationships within a local context are captured. However, the research on the inter-document relationships and the more complicated in-document knowledge correlations are still under-studied.

Symbolic Dialogue Management. Knowledge graphs are commonly used in dialogue management, such as symbolic dialogue transition graph (Xu et al., 2020a) and common sense (Zhou et al., 2018) graph. Most of these methods (Xu et al., 2020b,a) learn dialogue sequence by reinforcement learning(RL), which requires sophisticated reward design and well-predefined transition graph. RL based dialogue system model dialogue sequence on the hypothesis of Markov Decision Process (Bellman, 1957), which in fact is not true multi-turn transition. There are also some studies (Moon et al., 2019; Ma et al., 2021) select knowledge by walking on the graph or attending to the graph node (Zhou et al., 2018). However, they ignores the dialogue sequence while using the graph structure for transitions. Our method performs knowledge selection by sufficiently capturing the graph information and incorporating a multi-turn knowledge difference sequence learning.

Sequential Modeling in Dialogue Existing studies (Kim et al., 2019; Zhan et al., 2021b) on knowledge grounded conversations also addressed historical knowledge sequence in knowledge grounded conversations. (Kim et al., 2019) captured knowl-

edge sequence by a latent variable. (Zhan et al., 2021b) further proposed to learn more abstracted topic sequence to avoid the knowledge sparse problem and knowledge transition noise. However, both studies didn't consider the historical knowledge difference/change sequence and ignored to include the structural knowledge information.

3 Approach

3.1 Task Definition

The document grounded conversation task is defined as follows. Given the dialogue context $U_t = \{u_{t-1}, ..., u_{t-l}\}$ and grounding documents $\mathcal{D}_t$, where u_{t-1} represents the lastest (must be user turn) utterance and u_{t-l} means the earliest (maybe user/bot turn) utterance within the context length l. The grounding document $\mathcal{D}_t$ generally consists of multiple passages covering different dialogue topics, which are denoted as $\{p_1, p_2, ..., p_{|\mathcal{D}_t|}\}$. Each passage is composed of a topic phrase t_i and multiple knowledge sentences which is denotes as $\{k^i_1, k^i_2, ..., k^i_{|p_i|}\}$ where $|p_i|$ is the number of sentences in passage p_i. The target for this task is to select the most reasonable knowledge sentence from the grounding documents $\mathcal{D}_t$ and generates the response based on the selected knowledge.

3.2 Graph Construction

To fully use the structural information, we present a Co-Referential Multi-Document Graph(CoRM-DoG) to incorporate document inter-correlations and the in-document knowledge relations. Specifically, we construct one CoRM-DoG for each data sample, denoted as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$, $\mathcal{E}$ are nodes and edges respectively.

Nodes. As shown in Fig 1, the CoRM-DoG contains two types of nodes: topic nodes and knowledge nodes. Each topic node represents one of passage p_i from $\mathcal{D}_t$ while each knowledge node indicates one knowledge sentence k^i_j of passage p_i.

Edges. There are three types of edges in our CoRM-DoG: 1) relations between topics and knowledge; 2) relations between topics; 3) relations between knowledge under the same topic. Therefore, knowledge under different topics is not connected in our graph. For the first type of relation, we use the sentence order of knowledge k^i_j in its corresponding passage p_i as the edge type between the knowledge and the topic, denoted as `sent_n` edge. The remaining two types of relations are illustrated as follows.

3.2.1 Topic Relations

Human-to-human conversations may perform topic transitions following the commonsense relations between two topics or simple topic similarity to keep engaging dialogue, such as from *UK* to *London*(commonsense) or from *sci-fi movie* to *sci-fi novel*(similarity). Inspired by this, we introduce two types of relations between topics, modeling the above two types of topic transitions respectively.

Word Overlap. We simply employ the word overlap between two topics to measure their similarity. Specifically, we obtain the lemmas of topics phrases by spaCy[1] and judge whether the two topics have at least one identical lemma, so as to determine whether these two topics nodes have a `word_overlap` edge.

Commonsense. We found that the knowledge backend of the WoW (Dinan et al., 2019) dataset comes from the Wikipedia corpus, so we use the WikiData[2] to obtain commonsense relations between topics. We only collected relations for topics in the training set. In post-processing, we kept the high-frequency relation types and uniformly treat other low-frequency relation types as `other_relation`.

3.2.2 Knowledge Relations

We propose a variety of ways to express the relations between the knowledge sentences $\{k^i_1, k^i_2, ..., k^i_{|p_i|}\}$ under the same passage p_i and found that co-reference relations achieve the best performance. Accordingly, the CoRM-DoG employs co-reference relations as edges between knowledge nodes.

Partial Order. We hypothesize that the writing logic contained in the document is compatible with the development of dialogue under the same topic to some extent. So we simply connect each knowledge node with several knowledge nodes behind according to the original text order, and we call this `partial_order_relation` edge. This relation guides the learning of dialogue management by introducing the sequence information of knowledge in the original passage. We explored the effects of partial order with different hops in Section 4.4

Entity Link. Diverting the dialogue context to the knowledge that shares the same entity with previous knowledge is another reasonable obser-

[1] https://spacy.io/, MIT License

[2] https://www.wikidata.org/wiki/Wikidata:Main_Page

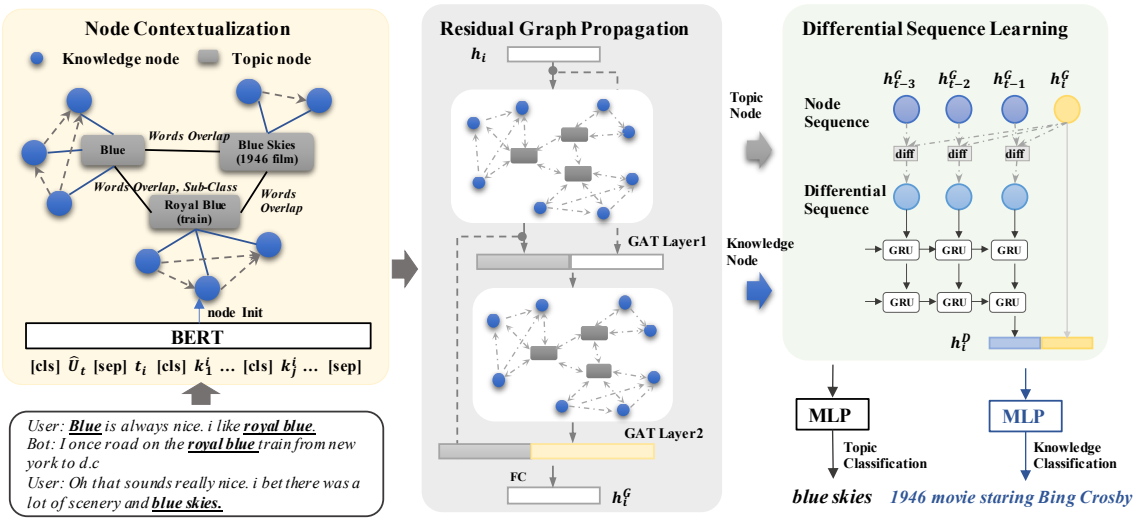


Figure 2: The architecture of proposed GraphDiffs. The left yellow part illustrates Node Contextualization with the BERT encoder. The middle gray part describes the graph information propagation through the proposed Residual Relation-aware Graph Attention(Res-RGAT). The temporal enhanced embedding from stacked GRU after Differential Sequence Learning which is shown in the right green part are further fed into MLP for multi-task(topics and knowledge selection) learning.

vation based on human-to-human conversations. We first extract the nouns from all knowledge sentences as entities by spaCy and then assign the `identical_entity_relation` edge to those knowledge with the same entity.

Co-reference Link. In addition to Partial Order and Entity Link, we also propose the `coreference_relation` edge. For each topic, the co-reference links (referring paths) within one passage p_i can be extracted by a co-reference resolution model[3]. For each co-reference link, every knowledge node on this link is connected to each other by `coreference_relation` edge.

3.3 Node Contextualization

In this part, we will introduce how to contextualize the topic and knowledge nodes in $\mathcal{G}$ by the U_t and $\mathcal{D}_t$. Following (Karpukhin et al., 2020; Cheng et al., 2020; Wu et al., 2021), we utilize BERT (Devlin et al., 2019) to obtain the representations for both topics and knowledge, which are used to initialize nodes in CoRM-DoG. In details, we concatenate the dialogue context U_t with each passage p_i separately and feed them into BERT encoder to get contextual representations. The concatenated input for one passage p_i is shown as follows,

$$[\text{cls}]\hat{U}_t[\text{sep}]t_i[\text{cls}]k^i_1...[\text{cls}]k^i_{|p_i|}[\text{sep}] \quad (1)$$

where $\hat{U}_t = [\text{usr}]u_{t-1}[\text{agt}]u_{t-2}...[\text{agt}]u_{t-l}$. The role symbols [usr] and [agt] are used to indicate utterance from the user turn or agent turn. We gather the first [cls] token embedding $\mathbf{t}_i$ to initialize the corresponding topic node of p_i. Similarly, the following [cls] token embedding $\{\mathbf{k}^i_1, \mathbf{k}^i_2, ..., \mathbf{k}^i_{|p_i|}\}$ are gathered and used to initialize the corresponding knowledge nodes under the same topic. Thus the process of contextualizing one specific passage p_i is formulated as:

$$\mathbf{t}_i, \mathbf{K}^i = \text{BERT}(U_t, p_i), i \in [1, |\mathcal{D}_t|] \quad (2)$$

where $\mathbf{t}_i, \mathbf{k}^i_j \in \mathbb{R}^d, \mathbf{K}^i = \{\mathbf{k}^i_j\}^{|p_i|}_{j=1}$. We can finally get contextualized node embedding as $\mathbf{H} = \{\mathbf{t}^i; \mathbf{K}^i\}^{|\mathcal{D}|}_{i=1} \in \mathbb{R}^{N\times d}$ where N is the total number of nodes, including both knowledge and topics.

3.4 Residual Graph Propagation

As described in Section 3.2, the CoRM-DoG contains many kinds of edges so we employ a look-up table $\mathbf{E} \in \mathbb{R}^{S\times d_e}$ to store the embedding of these edge types, here S is the number of edge types. Then getting the set of the edge embedding R for an CoRM-DoG $\mathcal{G}$ can be denoted as:

$$\mathbf{R} = \text{look\_up}(\mathbf{E}, \mathcal{E}) \quad (3)$$

As shown on the middle gray part in Fig 2, we use a relational variant of Graph Attention(GAT) layer (Veličković et al., 2017), denoted as RGAT layer in this work, to update the nodes representations by propagating information through the edges of $\mathcal{G}$. In specific, each node embedding is concatenated with the corresponding edge embedding

[3]https://github.com/huggingface/neuralcoref, MIT License

in one edge, which is then used to calculate the attention score.

Moreover, inspired by ResNet (He et al., 2016), we stack two RGAT layers through residual connection to propagate the information in multi-hop connections, which is named as Res-RGAT. Instead of using the sum operation in ResNet, we adopt a concatenation operation to avoid information loss. Besides, we use one Linear layer to transform the concatenated feature back to the same dimension with the input. The enhanced node representations are obtained after graph information propagation, which is formulated as:

$$\mathbf{H}^G = \text{Res-RGAT}(\mathbf{H}, \mathbf{R}, \mathcal{G}) \in \mathbb{R}^{N \times d} \quad (4)$$

3.5 Differential Sequence Learning

Sequential Modeling in dialogue also play a critical role in dialogue management (Kim et al., 2019; Zhan et al., 2021b). We propose a Differential Sequence Learning module to learn the sequential knowledge transition from dialogue context by GRU network (Cho et al., 2014), which is shown on the right green part in Fig 2. For knowledge or topic sequence appeared in previous agent turn (labels of the previous user turns are inaccessible in practice), we can collect their corresponding node representations from $\mathbf{H}^G$. We identically treat topic or knowledge sequences in two independent path, both can be denoted as $S = \{\mathbf{h}^G_{t-\tau}, ..., \mathbf{h}^G_{t-1}\}$. Then we obtain the differential sequence representations input with each item in S and each knowledge or topic node in $\mathbf{H}^G$ through difference evaluation function $\mathcal{F}$. Thus we can get one differential sequence for each node and there are N differential sequences in total, which is denoted as follows:

$$\{\mathcal{F}(\mathbf{h}^G_{t-\tau}, \mathbf{h}^G_i), ..., \mathcal{F}(\mathbf{h}^G_{t-1}, \mathbf{h}^G_i)\}^N_{i=1} \quad (5)$$

Inspired by (Chen et al., 2017; Zheng et al., 2020), we adopt difference and dot product operation as the difference evaluation function $\mathcal{F}$ to compute the dissimilarity of two vectors, denoted as $\mathcal{F}(\mathbf{a}, \mathbf{b}) = [\mathbf{a} - \mathbf{b}; \mathbf{a} \odot \mathbf{b}]$.

Each differential sequence is then fed into stacked GRU cells to learn and capture the sequential transition pattern. We concatenate the last hidden state of GRU with the node representation $\mathbf{h}^G_i$ to obtain differential sequence enhanced node representation $\mathbf{h}^D_i$, which is denoted as:

$$\mathbf{h}^D_i = [\text{GRU}(..., \mathcal{F}(\mathbf{h}^G_{t-1}, \mathbf{h}^G_i)); \mathbf{h}^G_i] \quad (6)$$

With N differential sequence we can get the Differential node representations $\mathbf{H}^D \in \mathbb{R}^{N \times 2d}$.

3.6 Training

In addition to the knowledge classification, topic classification is added as an auxiliary task to form a multi-task learning framework. We split the representation for topic nodes and knowledge nodes from $\mathbf{H}^D$ and get $\mathbf{H}^D_{tpc}$ and $\mathbf{H}^D_{knl}$ respectively. $\mathbf{H}^D_{tpc}$ is fed into a multi-layer perceptron (MLP) to obtain the topic selection scores. For the knowledge nodes, we include their corresponding topic node representation and the edge embedding between the knowledge node and the topic node to calculate the knowledge selection scores with MLP.

We also implement the history loss as an auxiliary objective function in our framework to further utilize the dialogue history information, which is the same as the recent work (Wu et al., 2021). We gather the context representations of each turn according to the embedding of role tokens [usr] and [agt] and calculate both history topics and knowledge loss with the labels of history turns.

In conclusion, the final objective function we adopt in this framework is formulated as follows:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{knl} + \mathcal{L}_{tpc} + \mathcal{L}_{hist} \\ \mathcal{L}_{hist} &= \frac{1}{2l} \sum_{hi=1}^{l} (\mathcal{L}^{hi}_{knl} + \mathcal{L}^{hi}_{tpc}) \end{aligned} \quad (7)$$

where l is a hyperparameter representing the longest context length. $\mathcal{L}_{knl}$ and $\mathcal{L}_{tpc}$ are knowledge loss and topics loss respectively. All the objective functions in $\mathcal{L}$ for classification are standard softmax cross-entropy.

4 Experiments

4.1 Experimental Settings

Datasets. We validate our method on two public benchmarks for the document grounded conversation, Wizard of Wikipedia (WoW) (Dinan et al., 2018) and Holl-E (Moghe et al., 2018). Dialogues in WoW are constructed based on the retrieved knowledge from Wikipedia, covering around 1300 topics and containing 18430/1948/965/968 dialogues for train/valid/test Seen/test Unseen sets. The test Seen set has topics overlap with the training data while topics in the test Unseen set are never seen in training set. Holl-E is another similar data set in the movie domain, including 7228/930/913 dialogues for train/valid/test sets. We use the same data setting as in (Kim et al., 2019).

Evaluation metrics. We mainly focus on the knowledge selection task for the document

grounded dialogue system, for which we use the knowledge and topic selection accuracy, denoted as ***Acc*** and ***T-Acc***, respectively. For the task of response generation with the dialogue context and selected knowledge, we calculate the overlap of the generated response and the ground-truth with the unigram-F1(***uF1***) and bigram-F1(***bF1***).

Baselines. We split baselines into three categories by their text encoder types. (i) **Non-Pretrained encoder**: Transformer+MemNet (Dinan et al., 2018) is the baseline released with the dataset WoW. DiffKS(RNN) (Zheng et al., 2020) incorporates knowledge difference feature in knowledge selection. (ii) **BERT encoder**: BERT+PoKS, a variant of PoKS with BERT (Devlin et al., 2019) encoder, learns knowledge selection by posterior knowledge distribution. SLKS (Kim et al., 2019) captures historical knowledge sequence with a latent variable. PIPM (Chen et al., 2020) improves SLKS by addressing the problem of missing posterior distribution in test phase. CoLV (Zhan et al., 2021a) includes two collaborative variables for knowledge selection and response generation. KnowledGPT (Zhao et al., 2020) optimizes knowledge grounded dialogue task by the pre-trained BERT encode and GPT-2 (Radford et al., 2019). (iii) **Passage-level BERT encoder**: DIALKI (Wu et al., 2021) encodes knowledge at passage level to capture knowledge relations, as we do in our GraphDiffs. As for response generation, the above mentioned methods adopted various generators so we uniformly replace their generators with the prompt-based generator Prefix-Tuning (Li and Liang, 2021), thus forming the baselines with "*" in Table 2.

4.2 Implementation Details.

We the BERT base model in all our experiments by using the Huggingface Transformers[4] (Wolf et al., 2020). We train the model with Adam (Kingma and Ba, 2015) optimizer and set the initial learning rate to 1e-5. A linear scheduler with a warm-up strategy in 5k steps is used here. The maximum history length l mentioned in Section 3.1 is set to 4 for WoW and 2 for Holl-E respectively to achieve the best performance. It takes around 5 and 10 epochs to achieve the reported performance by 4 nvidia V100 GPUs. We will release all the codes and all hyper-parameters settings for re-production.

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

Method	WoW Seen	WoW Unseen	Holl-E
	Acc	*Acc*	*Acc*
TMN	22.5	12.2	22.7
DiffKS(RNN)	25.6	18.6	33.5
BERT+PoKS	25.5	14.1	27.6
SLKS	26.8	18.3	29.2
PIPM	27.8	19.7	30.7
CoLV	30.1	18.9	32.7
DukeNet	26.4	19.6	30.0
KnowledGPT	28.0	25.4	-
DIALKI	32.9	35.5	-
GraphDiffs	**42.4**	**41.4**	**40.9**
-w/o Diff-Seq	40.8	39.5	39.7
-w/o Diff	40.9	40.1	40.1
-w/o Res-RGAT	35.5	36.5	39.5

Table 1: Knowledge Selection Results and Ablation Study on WoW and Holl-E.

Method	WoW Seen		WoW Unseen		Holl-E	
	uF1	*bF1*	*uF1*	*bF1*	*uF1*	*bF1*
SLKS(TM+Copy)	19.3	6.8	16.1	4.2	29.2	22.3
DukeNet(TM+Copy)	19.3	6.3	17.1	4.7	30.6	23.1
SLKS*	20.2	7.3	17.5	5.3	-	-
DiffKS*	21.5	7.6	20.0	6.3	30.7	23.9
KnowledGPT*	22.0	8.2	20.8	7.4	-	-
DIALKI*	22.0	8.0	22.2	8.1	-	-
GraphDiffs	**25.2**	**10.7**	**25.8**	**10.8**	**38.4**	**31.8**

Table 2: Response Generation on WoW and Holl-E. Methods with postfix "(TM+Copy)" use transformer generator with copy mechanism. The '*' after method name means the generator PrefixTuning (Li and Liang, 2021) is adopted. "-" means the method has no knowledge selection result on the dataset. *uF1* and *bF1* are unigram-F1 and bigram-F1.

4.3 Experimental Results

4.3.1 Metric-based Evaluation

Knowledge Selection. Knowledge selection results on both datasets are presented in Table 1. GraphDiffs significantly outperforms all other methods whether or not they employ BERT encoder. Compared with the recent best performance achieved by DIALKI, GraphDiffs improves by 9.5% and 5.9% and first achieves knowledge accuracy over 40% on the WoW Test Seen and Test Unseen sets. GraphDiffs also exceeds the TMN which is the baseline released with the WoW dataset by 19.9% on the Test Seen and 29.2% on the Test Unseen. For Holl-E, GraphDiffs still performs the best, with gains at least 7.4% in knowledge selection accuracy, compared with all previous state-of-the-art methods. The significant improvements on both datasets strongly prove that GraphDiffs can benefit from the graph modeling with differential sequence transition based on CoRM-DoG.

Response Generation. Table 2 shows the results of response generation on both WoW and Holl-E. We apply PrefixTunnig (Li and Liang, 2021) to generate responses given the concatenated dialogue context and selected knowledge as the input. The PrefixTunnig on BART (Lewis et al., 2020) obtains the comparable performance with fewer learnable parameters and extrapolates better to unseen topics than fine-tuning method. We conduct generation experiments with the same method for other systems, as illustrated in Table 2. GraphDiffs gets the best performance in the terms of all generation metrics compared with other methods. The reason is that GraphDiffs gains great improvement margins on knowledge selection performance compared to all baselines. For *uF1* and *bF1*, GraphDiffs exceeds the recent best DIAKI by $3.2\%/2.7\%$ on WoW Seen and $3.6\%/2.7\%$ on WoW Unseen.

4.3.2 Case Study

We give two visualized examples for the generated responses, as shown in Figure 3. The *Dialogue Context* rows are dialogue histories and the generated responses of different methods are listed in the *Response* row. We compare our GraphDiffs with DialKI and the Gold response. The first example performs topic change, from topic "hair loss" to "management of hair loss". GraphDiffs chose the right knowledge topic "management of hair loss" while DialKI repeated the knowledge mentioned in last conversation turn. The reason is that GraphDiffs referred to the `word_overlap` connection between "hair loss" and "management of hair loss". DialKI didn't consider the inter-topic relations, thus failed in this case. For the second example, the knowledge transition is within topic, denoted as **In-Topic** (knowledge in consecutive turns belonging to the same topic), our method successfully predicted the right knowledge due to the co-reference relation between these knowledge sentences in document "seattle". However, in the response generated by DialKI, even with passage-level knowledge correlations encoded, DialKI missed the longer dependencies, like from the 2nd sentence to the 6-th sentence in this case.

4.4 Analysis

Graph Structure Analysis. To analyze the effects of different graph structures we conduct several experiments on the WoW dataset. Three variants of CoRM-DoG are designed to study the relations among topics as follow: (1) *TP-w/o rela*: removing all relations; (2) *TP-w/o overlap*: removing the word overlap relation; (3) *TP-w/o wikigraph*: removing the commonsense relations. Three variants are designed to study the relations among knowledge as follow: (1) *KG-w/o rela*: removing all relations; (2) *KG-entity*: applying identical entity relation instead; (3) *KG-partial order*: using partial order relation instead. As shown in Table 3, for topic relations, removing both relations, performance drops more than removing one of them, indicating both word overlap relations and commonsense relations contribute to the knowledge selection accuracy. For knowledge relations analysis, we found the three variants of CoRM-DoG all lead to performance drop, which indicates that coreference relations(used in CoRM-DoG) are more appropriate for the conversation logic under the same topic. Surprisingly, *KG-w/o rela* achieves even higher results than *KG-entity*. One potential reason is that graph with entity relations introduces some wrong connections and brings noise to the model. For partial order relations, we explore the effects of different hops. Hop-k partial order relation means each knowledge node is connected with k knowledge nodes behind according to the sentence order. As shown in Fig 4, hop-2 partial relation performs the best. A hop that is too large or too small could cause information loss or introduce lots of wrong connections. While for coreference knowledge connections, more accurate links are preserved, thus achieving the best results. There are also some other kinds of relations we haven't tried in this work, such as discourse relations, we will save it for future work.

Method	**WoW Seen**			**WoW Unseen**		
	Acc	*T-Acc*	*In T-Acc*	*Acc*	*T-Acc*	*In T-Acc*
DIALKI	32.9	70.0	42.3	35.5	71.6	43.5
GraphDiffs	42.4	76.1	51.1	41.4	77.7	49.2
TP-w/o rela	42.1	74.0	50.6	39.8	75.2	47.2
TP-w/o overlap	42.4	75.9	51.2	40.9	77.7	48.1
TP-w/o wikigraph	42.3	75.9	50.9	41.1	77.5	48.8
KG-w/o rela	35.4	75.7	44.6	37.1	77.2	46.1
KG-entity	35.4	74.6	44.4	36.4	75.9	43.8
KG-partial order	36.6	75.9	45.7	37.1	76.8	45.5

Table 3: Graph Structure Comparison. *T-Acc* refers to topic selection accuracy. *In T-Acc* is In-topic knowledge selection accuracy. As the strongest baseline on WoW, DIALKI also report the accuracy of topic selection.

Ablation Study. To study the impact of different parts in GraphDiffs we conduct three additional experiments. *w/o Diff-Seq* means removing the Differential Sequence Learning while *w/o Diff*

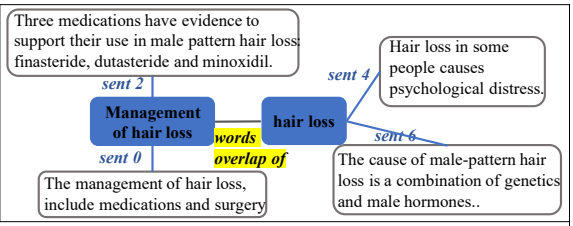
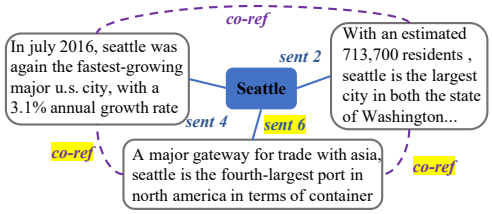
Dialogue Context	***User***: Hair loss, I'm going bald! ..... ***Bot***: There are surgeries and medications that can help .... **[management of hair loss', '0']** ***User***: What is the top option! I need help! ***Bot***: Typically, the reasons for hair loss in men has to do with genetics and male hormones.... **['hair loss', '6']** ***User***: Thanks for all your help/info today .	
Responses	***GD-SKT***: You are welcome. I would recommend taking three medications to help with male pattern hair loss: finasteride, dutasteride and minoxidil. ***DialKI***: well, I hope you have a great day! I know that hair loss can be caused by both genetic and environmental factors. ***Gold*** : The three types of medications that seem to help the most are finasteride, dutasteride, and minoxidil. I hope this information will help you! **['management of hair loss', '2']**	
Dialogue Context	***User***: Seattle, I have never been to Seattle but I would love to visit. ***Bot***: I would too. I read it's the fastest growing city in 2016, with a 3.1 annual growth rate. **['seattle', '4']** ***User***: wow, I bet it is really busy and crowded. ***Bot***: I believe it said there were over 700,000 residents just in Seattle and it is the largest city in Washington also. **['seattle', '2']** ***User***: I would hate to drive there. I'm assuming people use a lot of public transportation.	
Responses	***GD-SKT***: I'm not sure, but it is a major gateway for trade with Asia and the fourth largest port in north America. ***DialKI***: I'm not sure but I do know that it is the most populous city in the United States. ***Gold***: me too. I hate waiting in traffic. it's a major trade route with Asia. It has the fourth largest port in north America in terms of container shipping. **['seattle', '6']**	

Figure 3: Two generation examples from WoW. The bold words in "[]" indicate the knowledge. For example, **["hair loss", 6]** represents the 6-th knowledge sentence in the document with topic **"hair loss"**. Our method chose the right knowledge for both examples compare to DialKI owing to the well-designed graph structure.

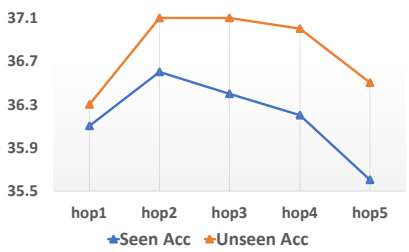


Figure 4: Knowledge accuracy for partial order with different hops.

uses the normal sequence instead of the differential sequence. *w/o Res-RGAT* removes the Residual Graph Propagation. The dramatically degraded performance of *w/o Res-RGAT* indicates GraphDiffs largely benefits from the proposed CoRM-DoG which guides the model to learn the effective knowledge transitions through graph structure. On the other hand, *w/o diff-seq* and *w/o Diff* both show that Differential Sequence Learning contributes to the best performance for taking the sequential knowledge transition into consideration.

In-Topic Knowledge Selection. Besides **T-Acc** and **Acc**, we also propose a new metric **In T-Acc** to evaluate the knowledge selection under the same topic. We extract samples that have the same topic as the last turn to calculate the accuracy. The results are shown in Table 3. The variants of topic relations achieve comparable performance, which reveals topic relations have little effect on knowledge selection under topic-invariant. The best results of GraphDiffs with the coreference relations again confirm the advantage of applying the novel coreference link in passage text as in-document knowledge relations. *KG-w/o rela* also performs better than *KG-entity* under topic-invariant samples, which indicates that identical entity relations are so dense that noise edges are included to cause the performance degradation.

5 Conclusion

We are the first to introduce the document graph for dialogue flows of the document grounded dialogue system. A highly effective document graph named Co-Referential Multi-Document Graph(CoRM-DoG) is proposed, which includes both inter-document relations and the co-referential intra-document connections. To fully use the document graph and adapt it to dialogue, we develop a novel Graph Modeling with Differential Sequence(GraphDiffs) method to simultaneously exploit structural knowledge and dialogue-specific sequence information. GraphDiffs based on CoRM-DoG has been empirically shown to make great progress on the knowledge selection task for document grounded conversation.

6 Ethical Impact

Our work aims to address the core knowledge selection problem in document grounded conversation. We encourage future work to propose a more meaningful and promising idea based on our strong baseline in document grounded conversation. We believe that document-grounded dialogue technology has broad application prospects in open-domain dialogue, emotional escort robots, intelligent assistants, etc. Knowledge selection also plays a significant role in dialogue management of multi-turn dialogue. However, more advanced dialogue knowledge selection and localization techniques tend to enable bots to select harmful content on the Internet and generate inappropriate or biased responses to user. All datasets we used in this work were privacy filtered and content moderated by the dataset authors (Dinan et al., 2019; Moghe et al., 2018).

References

Richard Bellman. 1957. A markovian decision process. *Journal of mathematics and mechanics*, 6(5):679–684.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668.

Xiuyi Chen, Fandong Meng, Peng Li, Feilong Chen, Shuang Xu, Bo Xu, and Jie Zhou. 2020. Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3426–3437, Online. Association for Computational Linguistics.

Hao Cheng, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2020. Probabilistic assumptions matter: Improved models for distantly-supervised document-level question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5657–5667, Online. Association for Computational Linguistics.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

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

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Song Feng, Siva Reddy, Malihe Alikhani, He He, Yangfeng Ji, Mohit Iyyer, and Zhou Yu, editors. 2021. *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*. Association for Computational Linguistics, Online.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2019. Sequential latent knowledge selection for knowledge-grounded dialogue. In *International Conference on Learning Representations*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *IJCAI International Joint Conference on Artificial Intelligence*, page 5081.

Wenchang Ma, Ryuichi Takanobu, and Minlie Huang. 2021. CR-walker: Tree-structured graph reasoning and dialog acts for conversational recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1839–1851, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chuan Meng, Pengjie Ren, Zhumin Chen, Weiwei Sun, Zhaochun Ren, Zhaopeng Tu, and Maarten de Rijke. 2020. Dukenet: A dual knowledge interaction network for knowledge-grounded conversation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1151–1160.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2332, Brussels, Belgium. 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 Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.

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.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zeqiu Wu, Bo-Ru Lu, Hannaneh Hajishirzi, and Mari Ostendorf. 2021. DIALKI: Knowledge identification in conversational systems through dialogue-document contextualization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1852–1863, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Jun Xu, Haifeng Wang, Zhengyu Niu, Hua Wu, and Wanxiang Che. 2020b. Knowledge graph grounded goal planning for open-domain conversation generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9338–9345.

Haolan Zhan, Lei Shen, Hongshen Chen, and Hainan Zhang. 2021a. CoLV: A collaborative latent variable model for knowledge-grounded dialogue generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2250–2261, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haolan Zhan, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Yongjun Bao, and Yanyan Lan. 2021b. Augmenting knowledge-grounded conversations with sequential knowledge transition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5621–5630, Online. Association for Computational Linguistics.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online. Association for Computational Linguistics.

Chujie Zheng, Yunbo Cao, Daxin Jiang, and Minlie Huang. 2020. Difference-aware knowledge selection for knowledge-grounded conversation generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 115–125, Online. Association for Computational Linguistics.

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

A Implementation Details

We set the maximum lengths of model input to 512, which is also the longest input limit for the BERT model, in order to fit the longer passage text as much as possible on both datasets. We employ a Linear layer to transform the output features of BERT from 768 to 320 to reduce memory usage. The edge embedding size is set to 64. The hidden size and headers of Res-RGAT are 1024 and 8 respectively while the alpha value of Graph Attention Network is 0.2. We utilize a unidirectional stacked GRU model for Differential Sequential Learning, the number of GRU layers is 2.

For response generation, we apply PrefixTunning (Li and Liang, 2021) on BART (Lewis et al., 2020) large model to learn the responses generation model based on the knowledge selection results from the previous stage. We use the prefix length 200 and the hidden dimension of 800 for all the methods using PrefixTuning generator. The PrefixTuning generator takes about 4 hours and 30 epoch to become converged during training on 4 GTX3090 24G GPUs, which is much faster and more resource saving than fine-tuning BART large.

B Graph Construction Details

WoW. There are more than 130k different documents from Wikipedia in WoW training set. We keep 350 high-frequency relations from the Wiki knowledge graph, covering these 130k documents. The top-10 wiki relations with corresponding frequency are shown as follows:

1. ('subclass of', 27015)
2. ('facet of', 11381)
3. ('sport', 10646)
4. ('performer', 9482)
5. ('part of', 6892)
6. ('manufacturer', 5742)
7. ('instance of', 5551)
8. ('history of topic', 5517)
9. ('has part', 5445)
10. ('follows', 5077)

As shown in Table 4, we found the `word_overlap` edge is more dense than the `wiki_relations` edge for the topics relations as the average number of `wiki_relations` in one sample is smaller, which is shown in the following Table. While for knowledge relations, the `coreference_relation` has much less average number of relations in one sample than other two types relations, which again proves that `coreference_relation` with more accurate knowledge relations lead to better knowledge selection results without introducing wrong structures information to GraphDiffs framework.

	Topic Relations		**Knowledge relations**		
	WordOverlap	*WikiGraph*	*Partial*	*EntityLink*	*Coreference*
Freq	8.11	2.89	61.18	87.52	15.90

Table 4: Average number of different kinds of relations in one sample on the WoW training set.

Holl-E. Different from WoW, each sample of Holl-E has only one topic, which is the name of the movie in this session of conversation. There are four types of information for each movie in Holl-E, which are plots, comments, reviews, and table information. So we simply divide all the knowledge sentences of each movie into four topics. As the absence of common sense relations of such topics in Holl-E, we count the co-occurrence relationship of all topics in the training set as the relations between topics in Holl-E. The relations between knowledge are as same as the WoW, using coreference relations in passage text. The relations between knowledge and topics are sentence order of knowledge sentence in the original text, which is also used in WoW.