Multi-Grained Conversational Graph Network for Retrieval-based Dialogue Systems

Quan Tu,[1] **Chongyang Tao,**[2] **Rui Yan**[1]
[1] Gaoling School of Artificial Intelligence, Renmin University of China, [2]Microsoft
{quantu, ruiyan}@ruc.edu.cn, chongyangtao@gmail.com

Abstract

Retrieval-based dialogue agents aim at selecting a proper response according to multi-turn conversational history. Existing methods have achieved great progress in terms of retrieval accuracy on benchmarks with pre-trained language models. However, these methods simply concatenate all turns in the dialogue history as the input, ignoring the dialogue dependency and structural information between the utterances. Besides, they usually reason the relationship of the context-response pair at a single level of abstraction (e.g., utterance level), which can not comprehensively capture the fine-grained relation between the context and response. In this paper, we present the multi-grained conversational graph network (MCGN) that considers multiple levels of abstraction from dialogue histories and semantic dependencies within multi-turn dialogues for addressing. Evaluation results on two benchmarks indicate that the proposed multi-grained conversational graph network is helpful for dialogue context understanding and can bring consistent and significant improvement over the state-of-the-art methods.

Keywords: retrieval-based dialogue, conversational graph, multi-grain

1. Introduction

Encouraged by the applications in virtual assistants such as Amazon Alexa and social chatbots such as Microsoft XiaoIce, there is a surge of interest in building a dialogue system that can conduct natural conversations with humans on open domain topics (Vinyals and Le, 2015; Liu et al., 2021). Existing implementations of such systems either select a proper response from existing conversations with information retrieval techniques (Lowe et al., 2015; Zhou et al., 2018; Humeau et al., 2020), or synthesize a response with natural language generation techniques (Vinyals and Le, 2015; Madotto et al., 2020; Roller et al., 2021; Liu et al., 2021). In this work, we study retrieval-based methods for open-domain dialogues, since retrieval models are superior in terms of response fluency and response informativeness, and thus play an important role in industrial products.

Real-world dialogues usually contain multiple utterances, where a retrieval model should select the most proper response by measuring the matching degree between multi-turn dialogue context and a bundle of response candidates. The key problem is how to make better use of multi-turn context information. Currently, there are two lines of research to represent the multi-turn dialogue context. One is to model each turn of utterance individually first and then aggregate a sequence of utterance-response matching features to get a final score (Wu et al., 2017; Zhou et al., 2018; Gu et al., 2019), which are known as the representation-matching-aggregation paradigm. The other line is to concatenate all turns of utterances into a long sequence first and make them fully interact with each other by RNNs (Lowe et al., 2015; Zhou et al., 2016; Chen and Wang, 2019) or transformer layers (Humeau et al., 2020; Whang et al., 2020; Gu et al., 2020).

Although existing mainstream methods have achieved impressive results in context modeling and response prediction, there are still two major limitations of these approaches. Firstly, existing models that fully concatenate all utterances or independently represent the information of each dialogue turn ignore the dialogue dependency and structure information between the utterances, which may lead to sub-optimal context representations and response matching features. Previous studies (Jia et al., 2020) have demonstrated that the semantic dependency among utterances is crucial for multi-turn response selection. Thus, how to model the dependencies in utterances remains a challenging problem for context understanding. Second, current response selection models usually represent the dialogue context and the response candidates and reason their relationship at single levels of abstraction (e.g., utterance level). We argue that explicitly representing multiple levels of abstraction (such as word-level and utterance-level) should make it easier for models to remember and reason over long-term context, and to predict appropriate responses with compositional structure. For example, it have been validated that keyword-levels of abstraction is effective in modeling the text sequence on dialogue generation (Serban et al., 2017a) and story generation (Chen et al., 2021).

To overcome the weaknesses of existing models and strive for better modeling multi-turn dialogues,

Corresponding author: Rui Yan (ruiyan@ruc.edu.cn)

LREC-COLING 2024, pages 11756–11765

we present the multi-grained conversational graph network (MCGN) that considers multiple levels of abstraction from dialogue histories and semantic dependencies within multi-turn dialogues to reason the relationship of the response candidates. More specifically, our MCGN consists of two graph-based branches to model the semantic flow of dialogue. In the first branch, we extract word information in utterance sequences to construct a word-level graph for the given multi-turn dialogue, so as to model the fine-grained topic transition dynamics. In the second branch, to model the semantic coherence of dialogue turns, we construct a discourse-level graph based on all utterances in dialogue and utilize the chronological order to indicate the weight of the edges. Our model employs the recent pre-trained language models (PLMs) (Devlin et al., 2019) to encode all inputs for better representing both dialogue context and candidate responses, and exploits the advantage of graph attention network (GAT) (Veličković et al., 2018) in properly aggregating information from other utterances over the two constructed graphs. By this means, our model can not only identify relevant contexts scattered across utterances but also capture more accurate semantic transition information with compositional graph structure.

We conduct experiments with two benchmarks, including DailyDialog Corpus (Li et al., 2017) and PersonaChat (Zhang et al., 2018a). On both benchmarks, the model is required to select the most appropriate response from a bundle of candidates. Evaluation results show that our proposed MCGN is significantly better than all state-of-the-art models on both datasets. Compared with the previous state-of-the-art methods, our model achieves 1.8% absolute improvement in terms of hits@1 (namely $R_{10}@1$) for the DailyDialog and 1.6% absolute improvement for the PersonaChat. In summary, our contributions are three-fold as follows:

- We propose a multi-grained conversational graph network (MCGN) for retrieval-based dialogue.
- We consider multiple levels of abstraction of dialogue including the word- and discourse-level for relationship representation.
- We achieve new state-of-the-art results on two benchmark datasets of open-domain dialogue.

2. Related Works

2.1. Retrieval-based Dialogue

Early work for retrieval-based dialogue systems studies single-turn response selection where the input of a matching model is a message-response pair (Wang et al., 2013; Ji et al., 2014; Wang et al., 2015). Recently, more attention is drawn to context-response matching for multi-turn response selection. Representative methods include the dual LSTM model (Lowe et al., 2015), the multi-view matching model (Zhou et al., 2016), the sequential matching network (SMN) (Wu et al., 2017), the deep attention matching network (DAM) (Zhou et al., 2018), and the multi-hop selector network (MSN) (Yuan et al., 2019).

Recently, pre-trained language models (Devlin et al., 2019; Liu et al., 2020) have shown significant benefits for various downstream natural language processing tasks, and many researchers have tried to exploit them on response selection. Vig and Ramea (2019) utilize BERT to represent each utterance-response pair and fuse these representations to calculate the matching score; Whang et al. (2020) treat the context as a long sequence and conduct context-response matching with BERT. Besides, the model also introduces the next utterance prediction and masked language model tasks borrowed from BERT to incorporate in-domain knowledge for the matching model; Gu et al. (2020) heuristically incorporate speaker-aware embeddings into BERT to promote the capability of context understanding in multi-turn dialogues.

2.2. Multi-turn Context Modeling

As a crucial problem in dialogue systems, multi-turn context modeling and understanding has raised great interest in the past few years. Especially for generation-based methods, various models adopt hierarchical encoder-decoder framework to model sequential context sentences (Serban et al., 2016, 2017c; Chen et al., 2018). Serban et al. (2017b) propose a multi-resolution RNN for modeling sequential data at multiple language granularity. Zhang et al. (2019) present ReCoSa model where attention weights between each context and response representations are computed and used in the further decoding process. (Hu et al., 2019) and Li et al. (2021) generalize existing sequence-based models to graph-structured neural network for dialogue generation. In retrieval-based dialogue, Zhang et al. (2018b) empirically concatenate the last utterance to other turns, and then use gated self-attention to obtain a query-aware utterance representation. Yuan et al. (2019) utilize multi-hop selectors to select the useful information in dialogue history and then perform the matching with the filtered context. Wu et al. (2018) introduce the topic information to enrich the semantics of the context and the response candidate, and introduce extra matching channels.

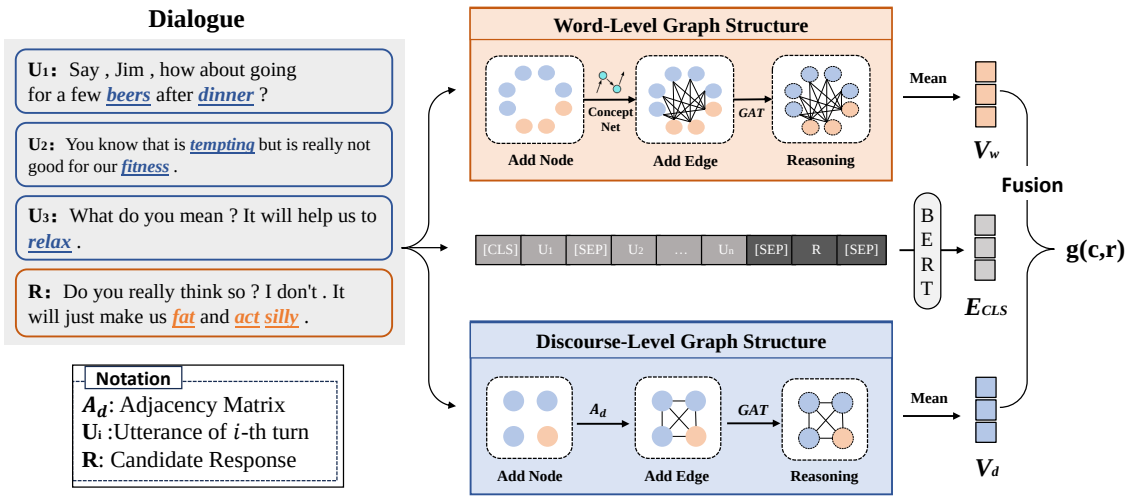

Figure 1: Overall architecture of our MCGN model.

3. Problem Formalization

Suppose that there is a multi-turn dialogue dataset $\mathcal{D} = \{c_i, r_i, y_i\}_{i=1}^{N}$, where $c_i = \{u_{i,1}, u_{i,2}, \ldots, u_{i,m_i}\}$ denotes a dialogue context with $u_{i,t}$ representing the utterance of the t-th turn, r_i denotes a response candidate, and $y_i \in \{0, 1\}$ denotes a label with $y_i = 1$ indicating that r_i is a proper response for c_i (otherwise, $y_i = 0$). Our task is to learn a context-response matching model $g(\cdot, \cdot)$ from $\mathcal{D}$ so that for any new context $c = \{u_1, u_2, \ldots, u_m\}$ and a response candidate r, $g(c, r) \in [0, 1]$ can compute the matching degree between c and r.

4. Model

4.1. Model Overview

The architecture of our proposed model is illustrated in Figure 1. The model has two graph-based branches that model the semantic flow of dialogue based on output representations generated by pre-trained language models. The first branch extracts topic-word information in utterance sequences to construct a word-level graph for the given multi-turn dialogue. This branch models the fine-grained topic transition dynamics. The second branch constructs a discourse-level graph based on all utterances in dialogue to capture the semantic coherence of dialogue turns. Our model uses the advanced graph attention network to aggregate information from other utterances over the two constructed graphs. Finally, our model computes the final matching score based on the aggregated features produced by both graphs and features produced by PLMs. By doing so, our model can identify relevant contexts scattered across utterances and capture more accurate semantic transition information with the multi-grained graph structure.

4.2. Encoding With BERT

We make use of the most recent pre-trained language models (PLMs) to encode all inputs, which helps in better representing both dialogue context and candidate responses. This serves as the foundation for the subsequent graph construction and response matching. Prior to diving into two semantic extraction branches, we introduce the details of multi-turn response selection with pre-trained language models (PLMs).

In particular, given a dialogue history $h = \{u_1, u_2, ..., u_m\}$ where u_i represents the i-th turn in the history, as well as a response candidate r, we concatenate all sequences as a single consecutive tokens sequence with special tokens. This is formulated as $x = \{[\texttt{CLS}], u_1, [\texttt{SEP}], \ldots, [\texttt{SEP}], u_m, [\texttt{SEP}], r, [\texttt{SEP}]\}$. Here, $[\texttt{CLS}]$ and $[\texttt{SEP}]$ are classification symbols and segment separation symbols respectively. For every token in x, BERT utilizes a summation of three kinds of embeddings, including WordPiece embedding (Wu et al., 2016), segment embedding, and position embedding. Subsequently, the embedding sequence of x is fed into BERT, giving us the contextualized embedding sequence $\{\mathbf{E}_{[\texttt{CLS}]}, \mathbf{E}_2, \ldots, \mathbf{E}_{l_x}\}$. $\mathbf{E}_{[\texttt{CLS}]}$ is an aggregated representation vector which contains the semantic interaction information between the context and response candidate.

4.3. Word-Level Graph Construction

In the first branch, we build a dialogue graph at the word level by utilizing the keyword information present in both the dialogue context $\boldsymbol{c}$ and response candidate $\boldsymbol{r}$. The goal is to capture the

coherence of the topic in a dialogue. The graph is denoted as $\mathcal{G}^w = (\mathcal{V}^w, \mathcal{E}^w, \mathcal{A}^w)$. Here, $\mathcal{V}^w$ represents a set of keyword nodes, $\mathcal{E}^w$ depicts a set of edges between topic words, and $\mathcal{A}^w$ represents the weight of the relational edge. We provide the details below.

Vertices. To establish the vertices in $\mathcal{G}^w$, we make use of a feature-based keyword extractor that combines both TF-IDF and POS (Huang et al., 2020) to extract the keywords of c and r. The keywords in c form the context-keyword vertices of $\mathcal{G}^w$, denoted as $\mathcal{V}_c^w = \{t_1, t_2, ..., t_p\}$. Similarly, the keywords in r form the response-keyword vertices of $\mathcal{G}^w$, denoted as $\mathcal{V}_r^w = \{t_{p+1}, t_{p+2}, ..., t_{p+q}\}$, where p and q are the numbers of keywords in the context c and the response r respectively. Therefore, $\mathcal{V}^w = \mathcal{V}_c^w \cup \mathcal{V}_r^w$. Once we have selected the vertices, we obtain vertex representations $\{\boldsymbol{h}_i^w\}_{i=1}^{p+q})$ by using mean-pooling of the token representation produced by PLMs.

Edges. We only take into account the edges between the context nodes $\mathcal{V}_c^w$ and the response nodes $\mathcal{V}_r^w$ because our objective is to predict the semantic coherence between the dialogue context and the response candidate. Moreover, we consider $\mathcal{G}^w$ as a weighted undirected graph and assign a weight to each edge of $\mathcal{G}^w$ by using the hop information in the ConceptNet (Speer and Havasi, 2013), referred to as hop-attention weights. Specifically, let the weighted adjacency matrix of $\mathcal{G}^w$ be denoted as $\boldsymbol{\mathcal{A}^w}$, then the hop-attention weight of the edge between the nodes $\mathcal{V}^w(i)$ and $\mathcal{V}^w(j)$ (i.e., $\boldsymbol{\mathcal{A}}_{ij}^w$) is determined by:

$$\boldsymbol{\mathcal{A}}_{ij}^w = \frac{1}{d\big(\mathcal{V}_c^w(i), \mathcal{V}_r^w(j)\big)}, \tag{1}$$

where $d(\cdot)$ represents the shortest path between $\mathcal{V}_c^w(i)$ and $\mathcal{V}_r^w(j)$ over the ConceptNet graph. The idea is to redefine the distances between keyword nodes so that the nodes that are far away from each other have low weight values. Following Rong et al. (2020), we randomly deactivate a certain number of edges from $\mathcal{G}^w$ at each training step and normalize the adjacency matrix $\boldsymbol{\mathcal{A}}^w$ to prevent over-smoothing. The process is defined as:

$$\bar{\boldsymbol{\mathcal{A}}}^w = (\boldsymbol{D}^w + \boldsymbol{I})^{-1/2}(\boldsymbol{\mathcal{A}}^w + \boldsymbol{I})(\boldsymbol{D}^w + \boldsymbol{I})^{-1/2}, \tag{2}$$

where $\bar{\boldsymbol{\mathcal{A}}}^w$ is the augmented normalized adjacency matrix, $\boldsymbol{D}^w$ is the corresponding degree matrix of $\boldsymbol{\mathcal{A}}^w$, and $\boldsymbol{I}$ is the identity matrix.

4.4. Discourse-Level Graph

I can definitely reword each sentence. Here's the result:

We create a discourse-level graph that utilizes all dialog utterances to capture the semantic coherence of the conversation, in addition to the word-level graph.

To represent a conversation with m utterances, we construct an undirected graph $\mathcal{G}^d = (\mathcal{V}^d, \mathcal{E}^d, \mathcal{A}^d)$, where $\mathcal{V}^d$ is the utterance nodes, $\mathcal{E}^d$ is the edges between words, and $\mathcal{A}^d$ is the weight of the relational edge.

We create the graph by following these steps:

Vertices: Each turn in the dialog history $c = \{u_1, \ldots, u_m\}$ and the response candidate r are treated as a vertex $\mathcal{V}^d(i)$. Thus, there are $m+1$ vertices in $\mathcal{G}^d$.

To initialize each vertex, we use BERT (Devlin et al., 2019) to encode each utterance in context and the response. We then take the representation of [CLS] as the utterance-level contextualized representation. The vertex representation of $\mathcal{G}^d$ is denoted as $\{\boldsymbol{h}_i^d)\}_{i=1}^{m+1}$.

Edges: We define the dialogue as an undirected complete graph where each vertex has an edge to all other vertices.

We utilize the chronological order to indicate the edge weight, since each utterance is contextually dependent on its adjacent utterances in a conversation.

For vertex $\mathcal{V}^d(j)$, we compute the weight of incoming edge $\mathcal{A}_{ij}^d$ as follows:

$$\boldsymbol{\mathcal{A}}_{ij}^d = \frac{1}{||i-j|| + 1} \tag{3}$$

To prevent over-smoothing, we perform the same operation as described in Equation (2) following Rong et al. (2020). We denote the smoothed weighted adjacency matrix of $\mathcal{G}^d$ as $\bar{\boldsymbol{\mathcal{A}}}^d$.

4.5. Graph Updating and Aggregation

We use both the word-level graph $\mathcal{G}^w$ and discourse-level graph $\mathcal{G}^d$ to reason through a graph attention network (GAT) (Veličković et al., 2018). This allows us to explicitly model the dynamics of topic and semantic transitions. In mathematical formalization, we don't differentiate between the two graphs using superscript notation. GAT takes all nodes as input and updates the node feature $\boldsymbol{h}_i$ based on its neighboring nodes in the graph. The node's aggregated representation $\boldsymbol{z}_i^{(l)}$ at layer l is formulated as follows for the node $\mathcal{V}(i)$:

$$\boldsymbol{z}_i^{(l)} = \texttt{LeakyReLU}\Big(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W}_\mathbf{g} \boldsymbol{h}_j^{(l)}\Big), \tag{4}$$

where $\boldsymbol{h}_i^{(0)} = \bar{\boldsymbol{h}}_i$, $\mathcal{N}_i$ is the neighboring nodes of $\mathcal{V}(i)$ in the dialogue graph, $\mathbf{W}_\mathbf{g} \in \mathbb{R}^{d \times d}$ is learnable

parameters, α_{ij} is the attention coefficient, which can be calculated by:

$$\begin{aligned} \alpha_{ij} &= \frac{\exp\left(e_{ij}\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(e_{ik}\right)}, \\ e_{ij} &= \bar{\boldsymbol{\mathcal{A}}}_{ij} * \left(\boldsymbol{a}_l^T\left[\boldsymbol{h}_i^{(l)};\boldsymbol{h}_j^{(l)}\right]\right), \end{aligned} \tag{5}$$

In the l-th layer, $\boldsymbol{a}_l \in \mathbb{R}^{2d}$ is a set of learnable parameters. It should be noted that the attention coefficients are scaled with the augmented normalized adjacency matrix $\bar{\boldsymbol{\mathcal{A}}}$, as shown in Equation (5). This is done to ensure that the network pays more attention to the nodes that are closer to v_i in the ConceptNet graph during aggregation.

Afterwards, the aggregated representation $\boldsymbol{z}_i^{(l)}$ is combined with the i^{th} node representation $\boldsymbol{h}_i^{(l)}$ to obtain the updated node representation $\boldsymbol{h}_i^{(l+1)}$.

$$\boldsymbol{h}_i^{(l+1)} = \texttt{ELU}\left(\mathbf{W}_a\boldsymbol{h}_i^{(l)} + \boldsymbol{z}_i^{(l)}\right), \tag{6}$$

where $\mathbf{W}_a \in \mathbb{R}^{d\times d}$ is the weight matrix to transform $\boldsymbol{h}_i^{(l)}$ and $\texttt{ELU}(\cdot)$ is an exponential linear unit (Clevert et al., 2016).

After performing the above graph updating and aggregation L times, we can obtained the overall graph representation $\boldsymbol{v}$ based on representations of all nodes in L-th layer, which is formulated as:

$$\boldsymbol{V} = f_{\texttt{MLP}}(\texttt{mean}(\{\boldsymbol{h}_i^{(L)}\})), \tag{7}$$

where $\boldsymbol{h}_i^{(L)}$ is the i-th node representation at the last layer, $mean$ represents mean pooling and $f_{\texttt{MLP}}(\cdot)$ is a fully-connected layer with a ELU activation. The above computation can be applied to the word-level graph and the discourse-level graph, yielding the word-level graph representation $\boldsymbol{V}_w$ and the discourse-level graph representation $\boldsymbol{V}_d$.

4.6. Computing Matching Score

Finally, we first concatenate the word-level graph representation, discourse-level graph representation and BERT features $E_{\texttt{[CLS]}}$, and then fed them into a non-linear layer to calculate the final matching score, which is formulated as:

$$g(c,r) = \sigma(\boldsymbol{W}[\boldsymbol{V}_w;\boldsymbol{V}_d;\mathbf{E}_{\texttt{CLS}}] + b) \tag{8}$$

where $\boldsymbol{W}$ and b is training parameters for response selection task, σ is a sigmoid function.

We learn $g(\cdot,\cdot)$ by minimizing cross entropy with $\mathcal{D}$. Let Θ denotes the parameters, then the learning objective of our model is:

$$\begin{aligned} \mathcal{J}_\Theta = -\sum_{i=1}^{N} & y_i\log(g(c_i,r_i)) + \\ & (1-y_i)\log(1-g(c_i,r_i)). \end{aligned} \tag{9}$$

All the learning objectives are optimized using back-propagation with Adam algorithm (Kingma and Ba, 2015).

5. Experiments

We test our MCGN on two benchmarks for multi-turn response selection in open-domain dialogue.

5.1. Dataset and Evaluation Metrics

DailyDialog (Li et al., 2017): The dataset is a multi-turn dialogue benchmark that covers various topics about our daily life. Specifically, the dataset includes $13118/1000/1000$ dialogues for training, validation, and testing respectively. To augment more data for training and testing, we reconstruct each set by taking each consecutive five utterances as a sub-dialogue. In each sub-dialogue, the first four utterances are treated as dialogue history and the last turn is the gold response. The negative response is randomly selected from the rest of the subset. Therefore, the training set contains abound 40K context-response pairs with the ratio of positive examples and negative examples as $1{:}1$. Both the validation set and test set contain 16K pairs with the ratio of positive examples and negative examples as $1{:}9$.

PersonaChat (Zhang et al., 2018a): The dataset is a crowd-sourced dataset that consists of 8939 chit-chat dialogues for training, 1000 for validation, and 968 for testing. Positive responses are true responses from humans and negative ones are randomly sampled from the dataset. The ratio between positive and negative responses is $1{:}1$ in the training set, and $1{:}9$ in the validation and testing sets. Following Wu et al. (2017), we also employ hits@k (equivalent to $\mathrm{R}_n@k$, $n = 10$), and mean reciprocal rank (MRR) as evaluation metrics. Table 1 give more details about two datasets.

why not test the model on UTC/Douban

5.2. Baselines

We compared our model with the following representative models.

- Dual-LSTM (Lowe et al., 2017): the model first concatenates all utterances in the context to form a single sequence, and then uses an LSTM to produce the representations for the context and response individually. Finally, the model calculates a matching score based on their representations.
- SMN (Wu et al., 2017): the model lets each utterance in the context interacts with the response candidate, and forms matching vectors through CNNs. The matching vectors of all pairs are then aggregated with an RNN to calculate a matching score.
- DAM (Zhou et al., 2018): the model performs matching in a similar way as SMN but context utterances and a response are represented

Aspects	DailyDialog			PersonaChat		
	Train	Dev	Test	Train	Dev	Test
# context-response pairs	40K	16K	16K	34K	10K	10K
# candidates per context	2	10	10	2	10	10
# positive candidates per context	1	1	1	1	1	1
Avg. # turns per dialogue	8.29	8.48	8.21	12.28	12.34	12.21
Avg. # words per dialogue	83.52	81.60	83.62	153.39	153.29	152.80

Table 1: Details of the DailyDialog and the PersonaChat dataset.

Model	DailyDialog				PersonaChat			
	hits@1	hits@2	hits@5	MRR	hits@1	hits@2	hits@5	MRR
Dual-LSTM (Lowe et al., 2015)	50.21	71.16	92.43	69.63	39.83	63.97	90.03	61.77
SMN (Wu et al., 2017)	55.06	74.06	94.12	71.12	45.53	67.30	90.82	63.77
ESIM (Chen and Wang, 2019)	60.44	78.56	95.88	75.03	48.09	69.10	91.33	65.25
DAM (Zhou et al., 2018)	63.13	79.50	95.06	76.49	50.76	69.81	91.39	66.18
IoI (Tao et al., 2019)	63.06	80.56	94.81	76.56	52.97	70.43	92.66	66.53
Bi-Enc (Humeau et al., 2020)	79.31	90.69	98.75	87.49	64.60	79.30	95.20	77.21
Poly-Enc (Humeau et al., 2020)	79.88	91.19	98.44	87.77	65.20	80.80	94.70	77.69
Cross-Enc (Humeau et al., 2020)	86.75	95.75	99.69	92.41	75.30	87.20	97.50	84.62
MCGN	**88.56**	96.06	99.31	**93.29**	**76.90**	**89.00**	97.40	**85.75**

Table 2: Results on DailyDialog and PersonaChat dataset. Scores in bold are statistically significantly better than the state-of-the-art with $p < 0.05$ according to t-test.

with stacked self-attention and cross-attention layers. The matching vectors are aggregated with a 3-D CNN as a matching score.

- ESIM (Chen and Wang, 2019): the model first concatenates all utterances in the context into a single sequence, and then employs ESIM structure derived from NLI for context-response matching.
- IOI (Tao et al., 2019): the model lets the context-response matching process goes deep by stacking multiple interaction blocks. The matching information within an utterance-response pair is extracted and flows along the chain of the blocks via representations.
- Bi-Encoder (Humeau et al., 2020): the model is similar to Dual-LSTM, but a pre-trained language model is utilized to acquire context and candidate representations individually.
- Poly-Encoder (Humeau et al., 2020): the model uses the BERT to encode context and candidate, respectively. The context is represented with multi-vectors instead of just one in Bi-encoder. Context representations are then aggregated into a vector with an attention mechanism and interact with response representation.
- Cross-Encoder (Humeau et al., 2020): the model concatenates all utterances in the context and the response candidate, and fed them into BERT to perform full interactions. The final matching score is calculated based on the aggregated feature.

5.3. Implementation Details

In our experiments, we encode context and the response candidate through the English uncased $\text{BERT}_{\text{base}}$. For each example, we limit the maximum length of the concatenated context and response to 150 and 50 respectively. Intuitively, the last tokens in the dialogue history and the previous tokens in the response candidate are more important, so we cut off the previous tokens for the context but do the cut-off in the reverse direction for the response candidate. We vary the layer of GAT (L) in $\{1, 2, 3, 4, 5\}$, and find that $L = 3$ is the best choice. We select the number of keywords in each utterance in $\{1, 3, 5, 7, 9\}$ and choose 3. We train our model using Adam optimizer (Kingma and Ba, 2015). The initial learning rate is 0.00003 and keeps decaying during training. We set dropout as 0.1 and batch size as 32. We use the validation set to fine-tune hyper-parameters, and report results on the test set.

5.4. Evaluation Results

Table 2 reports evaluation results of our MCGN as well as the baseline methods on the two data

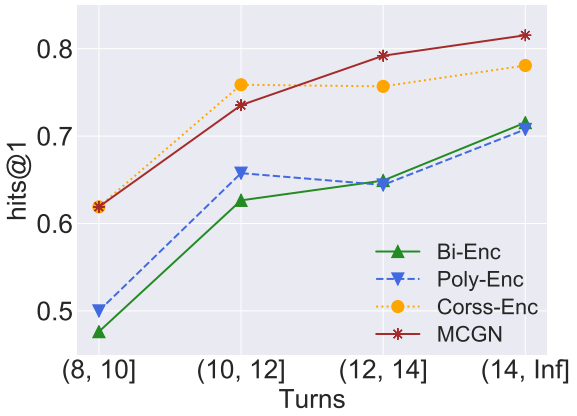

Figure 2: Performance of our MCGN and its variants across different turns of contexts.

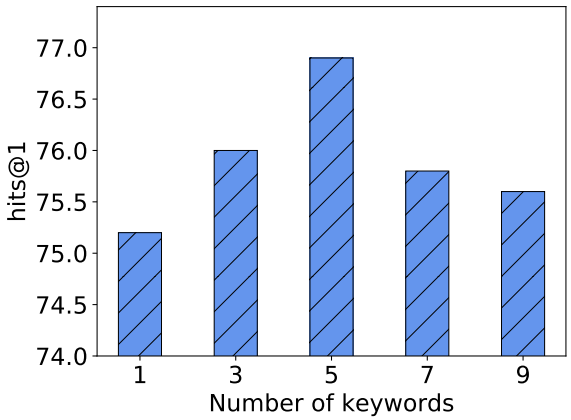

Figure 3: Performance of MCGN across different number of keywords.

sets. We can see that our proposed multi-grained conversational graph network can achieve better performance than all baseline models on both data sets, and improvement is statistically significant (t-test with p-value< 0.05) on most metrics. In particular, compared with the previous state-of-the-art method, our model achieves 1.8% absolute improvement in terms of hits@1 for the DailyDialog and 1.6% absolute improvement for the PersonaChat. Our multi-grained conversational graph network brings more obvious improvement on DailyDialog than that on PersonaChat. The difference may stem from that the conversations in DailyDialog are more natural and contain less topic shift than the PersonaChat, and therefore the structure modeling is more useful for DailyDialog.

Ablation Study. To investigate the impact of different graphs, we conducted a comprehensive ablation study. We keep the architecture of the matching model and remove each conversational graph individually from the model, and denote the model as "MCGN w/o. $\mathcal{T}$", where $\mathcal{T} \in \{\text{WG, DG}\}$ stand for word-level graph and discourse-level graph respectively. The detailed results are reported in Table 3. First of all, we find that two graphs are useful as removing any of them causes a performance drop on both datasets. Second, we can conclude that the discourse-level graph plays an important role in improving the response selection task. The reason might be that the discourse-level graph can encourage the model to consider the semantic coherence between the context and a response candidate, which is helpful to reason the relationship and acts as complementary to the word-level semantic interaction in PLMs and the word-level graph. It is also noted that removing any one leads to the more obvious decrease of the performance of response selection on DailyDialog, as the dialogues in PersonaChat include more topic shifts.

5.5. Discussions

Performance across different turns of context. To analyze how the performance of our proposed MCGN varies with different context lengths, we compare MCGN with Bi-Encoder, Poly-Encoder, and Cross-Encoder. Context length is measured by the number of turns in the dialogue history. Figure 2 shows how the performance of the four models varies across contexts with different lengths on PersonaChat. We can observe that the performance of all models increases monotonically as the context length increases. The results are rational since the model could capture more useful information for matching when more utterances are available in the context. Across the different lengths of the context, our MCGN can generally achieve better performance than Cross-Encoder as well as other baselines. It is worth noting that the performance of our MCGN is significantly better than other models for a long context. The results imply that our MCGN improves the capability to deal with long contexts with the multi-grained conversational graph structure.

Parameter Analysis. We first study how the number of keyword in the word-level graph influence the performance of our proposed model. Figure 3 shows how the performance of the models changes with respect to different numbers keyword on the test set of PersonaChat. We observe a similar trend for all models: they first increase monotonically until context length reaches 5, and then decreases when the number of keyword length keeps increasing. The reason might be that when only a few keywords are available in contexts, the model could not capture enough information for matching, but when keywords become enough, the noise will be brought to matching. Then we further analyze

Model	DailyDialog				PersonaChat			
	hits@1	hits@2	hits@5	MRR	hits@1	hits@2	hits@5	MRR
MCGN	88.56	96.06	99.31	93.29	76.90	89.00	97.40	85.75
MCGN w/o. DG	87.06	96.19	99.69	92.61	75.90	88.80	97.80	85.15
MCGN w/o. WG	87.56	96.44	99.44	92.88	76.00	88.60	97.80	85.21

Table 3: Ablation studies on DailyDialog and PersonaChat.

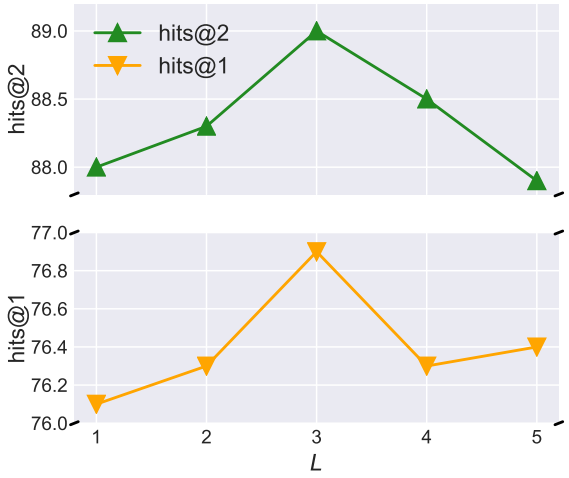

Figure 4: Performance of MCGN across different layers of GAT.

the effect of the number of layers (L) in GAT. Figure 4 illustrates how the performance of MCGN changed with respect to different ($L \in 1, 3, 5, 7, 9$) on the test set. It can be seen that the performance of our MCGN was obviously improved as L increased at the beginning, which shows the effectiveness of incorporating the contextual information between nodes with graph-based attention layers. Then, the performance was stable and dropped slightly. The reason might be that models begin to overfit due to a larger set of parameters.

6. Conclusion

In this paper, we consider the problem of multi-turn response selection in open-domain retrieval-based dialogue systems. Considering the fact that existing models usually ignore the dialogue dependency information between multi-turn utterances and reason the relationship of a context-response pair at single level of abstraction (e.g., utterance-level), we propose multi-grained conversational graph network (MCGN) for multi-turn response selection. The model considers multiple levels of abstraction of a dialogue and introduces two graphs to reason the relationship between the multi-turn context and the response candidates. We conduct experiments on two benchmarks and evaluation results show that the proposed multi-grained conversational graph Network are helpful for dialogue context understanding and can bring consistent and significant improvement over the state-of-the-art models. In the future, we would like to explore more dialogue structure information (such as event graph) to enhance the performance of the multi-turn response selection. We also want to valid the effectiveness of the proposed multi-grained conversational graph network (MCGN) on multi-turn response generation and conversational QA.

7. Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC Grant No. 62122089), Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, and Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China, the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China.

8. References

Hong Chen, Raphael Shu, Hiroya Takamura, and Hideki Nakayama. 2021. Graphplan: Story generation by planning with event graph. *arXiv preprint arXiv:2102.02977*.

Hongshen Chen, Zhaochun Ren, Jiliang Tang, Yihong Eric Zhao, and Dawei Yin. 2018. Hierarchical variational memory network for dialogue generation. In *Proceedings of the 2018 World Wide Web Conference*, pages 1653–1662.

Qian Chen and Wen Wang. 2019. Sequential matching model for end-to-end multi-turn response selection. In *ICASSP*, pages 7350–7354. IEEE.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2016. Fast and accurate deep network learning by exponential linear units (elus).

In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. In *CIKM*.

Jia-Chen Gu, Zhen-Hua Ling, and Quan Liu. 2019. Interactive matching network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2321–2324.

Wenpeng Hu, Zhangming Chan, Bing Liu, Dongyan Zhao, Jinwen Ma, and Rui Yan. 2019. Gsn: A graph-structured network for multi-party dialogues. *arXiv preprint arXiv:1905.13637*.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. Grade: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *ICLR*.

Zongcheng Ji, Zhengdong Lu, and Hang Li. 2014. An information retrieval approach to short text conversation. *arXiv preprint arXiv:1408.6988*.

Qi Jia, Yizhu Liu, Siyu Ren, Kenny Zhu, and Haifeng Tang. 2020. Multi-turn response selection using dialogue dependency relations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1911–1920, Online. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Yanran Li, Wenjie Li, and Zhitao Wang. 2021. Graph-structured context understanding for knowledge-grounded response generation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 1930–1934, New York, NY, USA. Association for Computing Machinery.

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*.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Roberta: A robustly optimized bert pretraining approach.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *SIGDIAL*, pages 285–294.

Ryan Lowe, Nissan Pow, Iulian Vlad Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. 2017. Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue & Discourse*, 8(1):31–65.

Andrea Madotto, Etsuko Ishii, Zhaojiang Lin, Sumanth Dathathri, and Pascale Fung. 2020. Plug-and-play conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2422–2433.

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

Yu Rong, Wen bing Huang, Tingyang Xu, and Junzhou Huang. 2020. Dropedge: Towards deep graph convolutional networks on node classification. In *International Conference on Learning Representations*.

Iulian Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron Courville. 2017a. Multiresolution recurrent neural networks: An application to dialogue response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, pages 3776–3783.

Iulian Vlad Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron Courville. 2017b. Multiresolution recurrent neural networks: An application to dialogue response generation. In *AAAI*, pages 3288–3294.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017c. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301.

Robert Speer and Catherine Havasi. 2013. Conceptnet 5: A large semantic network for relational knowledge. In *The People's Web Meets NLP*, pages 161–176. Springer.

Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues. In *ACL*, pages 1–11.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.

Jesse Vig and Kalai Ramea. 2019. Comparison of transfer-learning approaches for response selection in multi-turn conversations. In *Workshop on DSTC7*.

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

Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. 2013. A dataset for research on short-text conversations. In *EMNLP*, pages 935–945.

Mingxuan Wang, Zhengdong Lu, Hang Li, and Qun Liu. 2015. Syntax-based deep matching of short texts. In *AAAI*, pages 1354–1361.

Taesun Whang, Dongyub Lee, Chanhee Lee, Kisu Yang, Dongsuk Oh, and HeuiSeok Lim. 2020. An effective domain adaptive post-training method for bert in response selection. In *Proc. Interspeech 2020*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Yu Wu, Zhoujun Li, Wei Wu, and Ming Zhou. 2018. Response selection with topic clues for retrieval-based chatbots. *Neurocomputing*, 316:251–261.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Vancouver, Canada. Association for Computational Linguistics.

Chunyuan Yuan, Wei Zhou, Mingming Li, Shangwen Lv, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2019. Multi-hop selector network for multi-turn response selection in retrieval-based chatbots. In *EMNLP*, pages 111–120.

Hainan Zhang, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. 2019. Recosa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation. In *ACL*, pages 3721–3730.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018b. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3740–3752.

Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. 2016. Multi-view response selection for human-computer conversation. In *EMNLP*, pages 372–381.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *ACL*, volume 1, pages 1118–1127.