# CADGE: Context-Aware Dialogue Generation Enhanced with Graph-Structured Knowledge Aggregation

**Chen Tang[1], Hongbo Zhang[2], Tyler Loakman[2], Bohao Yang[1],**
**Stefan Goetze[2], Chenghua Lin[1]***

[1]Department of Computer Science, The University of Manchester, UK
[2]Department of Computer Science, The University of Sheffield, UK
{chen.tang,chenghua.lin}@manchester.ac.uk
bohao.yang-2@postgrad.manchester.ac.uk
{hzhang183,tcloakman1,s.goetze}@sheffield.ac.uk

## Abstract

Commonsense knowledge is crucial to many natural language processing tasks. Existing works usually incorporate graph knowledge with conventional graph neural networks (GNNs), resulting in a sequential pipeline that compartmentalizes the encoding processes for textual and graph-based knowledge. This compartmentalization does, however, not fully exploit the contextual interplay between these two types of input knowledge. In this paper, a novel context-aware graph-attention model (Context-aware GAT) is proposed, designed to effectively assimilate global features from relevant knowledge graphs through a context-enhanced knowledge aggregation mechanism. Specifically, the proposed framework employs an innovative approach to representation learning that harmonizes heterogeneous features by amalgamating flattened graph knowledge with text data. The hierarchical application of graph knowledge aggregation within connected subgraphs, complemented by contextual information, to bolster the generation of commonsense-driven dialogues is analyzed. Empirical results demonstrate that our framework outperforms conventional GNN-based language models in terms of performance. Both, automated and human evaluations affirm the significant performance enhancements achieved by our proposed model over the concept flow baseline.

## 1 Introduction

Open-domain dialogue generation has gained considerable traction in the field of natural language generation (Roller et al., 2021; Tang et al., 2023b). This task aims to develop chatbots with the capacity to engage in conversations across a broad spectrum of topics, thereby enabling a multitude of practical applications, including virtual assistants and well-being support systems (Abd Yusof et al., 2017;
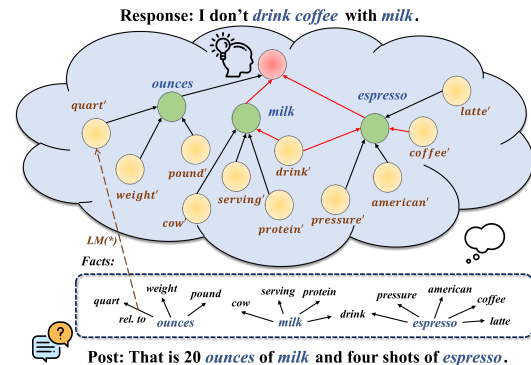
---

*corresponding author.



Figure 1: Illustration of the proposed model with an example. The retrieved facts are fed to the graph model, then the model learns the representations of concepts by aggregating the knowledge layer by layer. Finally, responses are generated with these aggregated features.

Wang et al., 2021; Tang et al., 2023c; Yang et al., 2024a). In recent years, there has been a surge of interest in leveraging large language models for dialogue generation (Zhang et al., 2019; Adiwardana et al., 2020; Roller et al., 2021; Tang et al., 2022b; Huang et al., 2022). These models, in general, exhibit an enhanced capacity to encapsulate knowledge within their networks as their model sizes increase. However, it is crucial to acknowledge a series of studies that have underscored the limitations of training on plain text corpora, where the knowledge structure is not explicitly represented during the learning process (Tang et al., 2022a; Yang et al., 2024b). Consequently, a key research question concerns how to better exploit and use external knowledge to improve the quality of generated responses, which has received increasing attention in recent research (Zhang et al., 2020; Yu et al., 2022; Wu et al., 2022; Tang et al., 2023a).

The knowledge incorporated into chatbots can be broadly divided into *structured* and *unstructured* forms. Prior work (Komeili et al., 2022; Ghazvininejad et al., 2018; Lian et al., 2019) has

achieved successful integration of unstructured knowledge (such as free-text content from web pages and knowledge statements) into the generated responses of chatbots. This typically involves encoding the most appropriate retrieved facts together with the conversation context or encoding multiple pieces of facts into a uniform representation before passing it to the decoder alongside the conversation history. Structured knowledge, on the other hand, usually takes the form of a graph. A range of neural models (Zhou et al., 2018; Yang et al., 2020; Lin et al., 2021) have been introduced to incorporate features from the retrieved graph-structured knowledge. For instance, the graph attention mechanism (Lotfi et al., 2021; Tuan et al., 2019; Zhou et al., 2018) has been widely used to embed knowledge graph features, and has been successful in aggregating sparse features into rich representations. With regard to language models, the rise of pre-trained models (Srivastava et al., 2021; Dong et al., 2019; Tang et al., 2024) has also substantially advanced the state-of-the-art (SOTA) in open-domain dialogue generation.

However, existing dialogue systems still face a number of challenges to effectively exploit commonsense knowledge (Xie et al., 2021). Since graph-structured knowledge and natural utterances have different representations, most prior work (Tuan et al., 2019; Zhou et al., 2018; Zhang et al., 2020) employed separate encoders to incorporate and leverage these heterogeneous features by concatenating their respective numeric vectors. However, since the separate encoders do not share low-dimensional representations, they may fail to fully account for the semantics of context contributed by given posts with additional external knowledge facts. In addition, existing frameworks directly conduct graph-attention-based encoding on retrieved facts from the knowledge base, which are isolated in separate sub-graphs. This strategy does not capture dependencies between sub-graphs nor between the graph knowledge and the context of the post, in turn making it hard for neural networks to fully capture the overall backgrounds from the inputs.

To address the aforementioned challenges, this paper proposes a novel graph-based framework to leverage knowledge contained in concept-related facts. In contrast to employing separate encoders to encode knowledge in the form of disparate knowledge graphs and text, we first transform the graph-structured representations into plain text, and leverage a pre-trained language model, UniLM (Dong

et al., 2019), to generate unified features for all inputs. Subsequently, to overcome inadequacies when capturing the context semantics provided by the given posts and retrieved knowledge facts, a novel, context-aware graph-based mechanism (Context-aware GAT) is proposed to incorporate the features from the post and the knowledge graph in the same learning process during hierarchical aggregation. The graph knowledge takes two steps (layers) before being aggregated into a condensed feature vector as the global features of given inputs. For each layer, the context embedding and the factual embedding are concatenated, and then graph attentions are computed for every sub-node. Finally, all representations are aggregated into the root node and fed to the decoder for response generation. This whole process is illustrated in Figure 1. We also note that our model can be easily extended to incorporate multi-hop knowledge. Experimental results show that our extended model can use multi-hop knowledge to further increase the informativeness of generated responses, and consequently yields considerable improvements over other dialogue systems that use multi-hop knowledge. The contributions of this work are summarised three-fold:

- We propose a novel framework[1], which is a successful exploration that leverages a unified language model for the heterogeneous inputs of graph knowledge and text, exploiting structured knowledge with context-aware subgraph aggregation to generate informative responses.
- We conduct a range of experiments, and the extensive automatic and human evaluation results demonstrate our model significantly outperforms existing baselines to generate a more appropriate and informative response with external graph knowledge.
- With extensive experiments, we investigate the advances and mechanisms of leveraging graph knowledge with our Context-aware GAT model. We also investigate the expansion of our model to accommodate multi-hop knowledge, and validate its effectiveness.

## 2 Related Work

Recently, much work has focused on augmenting dialogue systems with additional background knowledge. Such works can be divided into dia-

---

[1]Our code and datasets are accessible at `https://github.com/StevenZHB/CADGE`.

logue systems augmented with unstructured knowledge, and those augmented with structured knowledge. With unstructured knowledge, (Komeili et al., 2022) models web page information and feeds it into a language model. (Ghazvininejad et al., 2018) and (Lotfi et al., 2021) encode the filtered factual statements with a specific encoder and then pass them into the decoder along with context. (Lian et al., 2019) use context to aggregate knowledge statements and find that aggregated knowledge gives better results than filtered knowledge. Regarding structured knowledge, graph neural networks (Scarselli et al., 2009) are usually used to embed graph information to input into a language model. (Zhou et al., 2018) uses GRUs and two graph attention modules to select appropriate triples to incorporate into responses. In order to exploit the benefits of multi-hop knowledge, (Zhang et al., 2020) adds an attention mechanism in a similar way to filter the appropriate knowledge. Finally, (Tuan et al., 2019) proposes a model which selects the output from a sequence-to-sequence model and a multi-hop reasoning model at each time step.

Large language models such as UniLM (Dong et al., 2019), GPT-2 (Radford et al., 2019), and BART (Lewis et al., 2019) are widely used in open domain dialogue generation systems (Zeng et al., 2021). DialoGPT (Zhang et al., 2019) was pretrained on a dialogue dataset containing 147M conversations and is based on the autoregressive GPT-2 model, using a maximum mutual information (MMI) scoring function to address the low amount of information in the generated text. (Adiwardana et al., 2020) built a 2.6B-parameter Evolved Transformer architecture to model the relation between context-response pairs. To generate more informative responses, (Bao et al., 2019, 2020) use latent variables to model one-to-many relationships in context-response pairs. Finally, (Roller et al., 2021) use a retrieval model to retrieve candidate responses and then concatenates them to represent the context before inputting them into the transformer to generate the model. Please refer to Appendix A for more details of related work.

## 3 Methodology

We formulate our task as follows: The given inputs include a post $X = \{x_1, x_2, ..., x_n\}$ and a graph knowledge base $G = \{\tau_1, \tau_2, ..., \tau_k\}$, in which a fact is represented in the form of a triplet $\{h, r, t\}$ where $h$, $r$, and $t$ denote the head node, the relation,
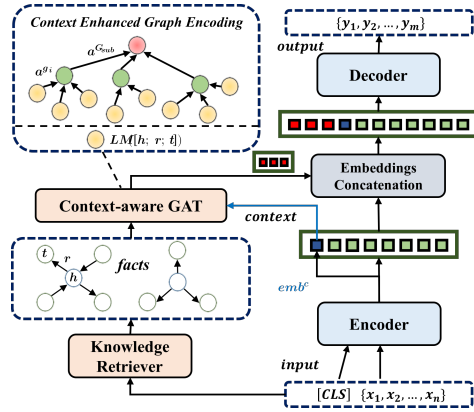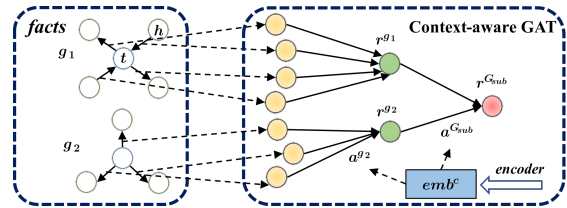


Figure 2: Overview of the proposed model.



Figure 3: The Context-aware GAT firstly transforms knowledge from facts into numeric vectors (in yellow). Through feature forwarding, the root nodes of each graph attentively read and aggregate all knowledge and become higher-level representations (from yellow to green, and then green to red).

and the tail node, respectively. The goal is to generate a response $Y = \{y_1, y_2, ..., y_m\}$ by modeling the conditional probability distribution $P(Y|X, G)$. Figure 2 gives an overview of our framework. The knowledge retrieval process is fundamentally implemented by word matching (concepts in ConceptNet are formatted in one-word) and rule filtering to collect knowledge triples (for more details please refer to (Zhou et al., 2018)).

### 3.1 Knowledge Representation

The 12-layer transformer blocks of UniLM (Dong et al., 2019) are split into two 6-layer parts - the encoder and decoder. When encoding the post's text, the language model of UniLM is informed of the high-level narrative structure using a classification label ([CLS]) to allow learning of the overall representation from $X$ as the context feature $emb^c$. For each recognised entity $ent_i$ in the post, relevant facts are retrieved from the knowledge base in the form of triples, and all retrieved facts can be considered as sub-graphs $g_i = \{\tau_1, \tau_2, ..., \tau_{N_{g_i}}\}$ in $G$. Each post usually results in several independent sub-graphs $G_{sub} = \{g_1, g_2, ..., g_{N_{G_{sub}}}\}$. In contrast to existing works that encode knowledge

in the form of disparate knowledge graphs and text, we propose to transform facts into text by directly concatenating them into a string, where they are then encoded with the embedding layer of UniLM:

$$E^{post} = LM([l_{[CLS]}; \{x_1, ...\}])) \quad (1)$$
$$= \{emb^c, emb_1...\} \quad (2)$$
$$f_e(h, r, t) = LM_{emb}([h; r; t]) \quad (3)$$
$$E^\tau = f_e(h, r, t) \quad \text{s.t.} \{h, r, t\} \in g_i \quad (4)$$

Operator $LM$ (abbr. of language model) denotes the encoder of UniLM, whilst $LM_{emb}$ denotes the embedding layer of UniLM, and $l_{[CLS]}$ denotes the "[CLS]" label.

### 3.2 Context-aware GAT

The overview of the proposed Context-aware GAT is as illustrated in Figure 3. The model learns the global graph features via translations operating on both the low-dimensional embeddings of the knowledge facts and the context contained in $emb^c$. To facilitate knowledge understanding and generation, we leverage a graph attention mechanism to aggregate knowledge representations layer by layer. With two layers of feature forward processing, we obtain the representation of the root node, $rt_{G_{sub}}$, as the aggregated feature for the whole graph, $G_{sub}$.

**First Forward Layer.** Our model firstly attends to the representations of facts $\tau \in g_i$ to compute graph attention and then aggregates features to the root node of each graph $rt^{g_i}$. The knowledge gradually updates the representations of root nodes step by step:

$$rt_t^{g_i} = \sum_{j=1}^{N_{g_i}} a_{tj}^{g_i} E_{tj}^\tau \quad (5)$$
$$a_{tj}^{g_i} = \frac{\exp(\beta_{tj}^{g_i})}{\sum_{j=1}^{N_{g_i}} \exp(\beta_{tj}^{g_i})} \quad (6)$$
$$\beta_j^{g_i} = W^{g_i}[E_{tj}^\tau; emb^c]^{\mathrm{T}} \quad (7)$$

where $t$ denotes the time step, $l_{pad}$ denotes the padding label to help initialize the root representations, and $W^{g_i}$ is a trainable parameter matrix.

**Second Forward Layer.** In analogy to the first forward layer, our model attends to the root nodes $rt^{g_i}$ represented for each sub-graph to attentively compute the final representation of the root node $rt^{G_{sub}}$, which stands for the overall features of all the retrieved sub-graphs:

$$rt_t^{G_{sub}} = \sum_{i=1}^{N_{G_{sub}}} a_{ti}^{G_{sub}}(rt_t^{g_i}) \quad (8)$$
$$a_{ti}^{G_{sub}} = \frac{\exp(\beta_{ti}^{G_{sub}})}{\sum_{i=1}^{N_{G_{sub}}} \exp(\beta_{tj}^{G_{sub}})} \quad (9)$$
$$\beta_i^{G_{sub}} = W^{G_{sub}}[rt_t^{g_i}; emb^c]^{\mathrm{T}} \quad (10)$$

### 3.3 Feature Aggregation and Decoding

After computing a representation for the root node, features from the post and retrieved knowledge are concatenated, and the decoder is employed to predict tokens $Y$ as the output response:

$$V = [rt^{G_{sub}}; E^{post}] \quad (11)$$
$$H = \text{Decoder}(V) \quad (12)$$
$$P(Y|X) = \text{softmax}(HW) \quad (13)$$

where $V$ denotes the aggregated features fed to the decoder, $H$ denotes the hidden states of the decoder used to predict the probability distribution of output tokens $P(Y|X)$, and $W$ is a trainable parameter.

### 3.4 Loss Function

**Auxiliary Entity Selection Task.** To better support representation learning, the entity selection task is introduced as an auxiliary task when training the proposed generative system. For each input post, the datasets contain corresponding annotations of knowledge triples $\Gamma = \{\tau_1', ..., \tau_{N_\Gamma}'\}$ from the knowledge base. These annotations can be considered as the ground truth of the knowledge paired with the post. The neural model is forced to select the ground-truth triples from all retrieved knowledge $G_{sub}$. As Figure 3 shows, each yellow node represents a knowledge triplet $\tau$, and each green node represents the root node $rt^{g_i}$. All yellow and green nodes have been labeled by checking if they are annotated as the ground truth. For instance, if $\tau_j \in \Gamma$ then the probability of $\tau_j$ denoted as $p_{es}(\tau_j|X)$ should be 1, and 0 otherwise. For the sub-graph root node (the green node in Figure 3), if $\tau_j \in g_i$ is the truth, then $p_{es}(rt^{g_i}|X)$ should be 1, and 0 otherwise. The probability distribution is modelled as follows:

$$p_{es}(\tau_j|X) = \text{softmax}(E_j^\tau W^{p_\tau}) \quad (14)$$
$$p_{es}(rt^{g_i}|X) = \text{softmax}((rt^{g_i})W^{p_{g_i}}) \quad (15)$$

where $es$ denotes entity selection, and $W$ denotes the trainable parameters.

**Overall Loss Function.** The loss function includes parts of the text prediction task and entity selection task, and is computed with cross entropy:

$$\mathcal{L}_{lm} = -\frac{1}{N} \sum_{n=1}^{N} \log P(Y|X) \tag{16}$$

$$\mathcal{L}_{es}^{\tau} = -\sum_{n=1}^{N} \sum_{j=1}^{N_\tau} s_j^{\tau} \log(p_{es}(\tau_j|X)) \tag{17}$$

$$\mathcal{L}_{es}^{g} = -\sum_{n=1}^{N} \sum_{i=1}^{N_{G_{sub}}} s_g^{\tau} \log(p_{es}(rt^{g_i}|X)) \tag{18}$$

$$\mathcal{L}_{overall} = \mathcal{L}_{lm} + \lambda_1 \mathcal{L}_{es}^{\tau} + \lambda_2 \mathcal{L}_{es}^{g} \tag{19}$$

where $N$ denotes the total amount of test data. $\lambda_1$ and $\lambda_2$ denotes the scale factors. $\mathcal{L}_{es}^{\tau}$ and $\mathcal{L}_{es}^{g}$ denote the loss of entity selections on the root nodes for facts $\tau$ and $g_i$, respectively. $\lambda_1$ and $\lambda_2$ are set to 1 in the following experiments.

### 3.5 Expansion for Multi-hop Knowledge

We also consider extending our model to incorporate multi-hop knowledge, which might give a further performance boost. Specifically, we extract the two-hop knowledge for all one-hop entities and use the same method to build a graph of two-hop knowledge. As the aggregation of two-hop knowledge needs to be related to one-hop knowledge, we use the one-hop knowledge aggregation representation $rt_{G_{sub}^{one}}$ in addition to the "[CLS]" label when aggregating two-hop knowledge. After passing through two layers of GAT, the root node of the two-hop knowledge graph ($G_{sub}^{two}$), $rt_{G_{sub}^{two}}$, which is treated as the aggregated features of the two-hop knowledge graph, is then concatenated with $rt_{G_{sub}^{one}}$ and input to the Decoder. The attention in the context-aware GAT for the two-hop knowledge graph is as follows:

$$a_{ti}^{G_{sub}^{two}} = \frac{\exp(\beta_{ti}^{G_{sub}^{two}})}{\sum_{i=1}^{N_{G_{sub}^{two}}} \exp(\beta_{tj}^{G_{sub}^{two}})} \tag{20}$$

$$\beta_i^{G_{sub}^{two}} = W^{G_{sub}^{two}}[rt_t^{g_i^{two}}; rt_{G_{sub}^{one}}; emb^c]^{\mathrm{T}} \tag{21}$$

$$a_{tj}^{g_i^{two}} = \frac{\exp(\beta_{tj}^{g_i^{two}})}{\sum_{j=1}^{N_{g_i^{two}}} \exp(\beta_{tj}^{g_i^{two}})} \tag{22}$$

$$\beta_j^{g_i^{two}} = W^{g_i^{two}}[E_{tj}^{\tau}; rt_{G_{sub}^{one}}; emb^c]^{\mathrm{T}} \tag{23}$$

The aggregated feature for the decoder is:

$$V_{mul} = [rt^{G_{sub}^{one}}; rt^{G_{sub}^{two}}; E^{post}] \tag{24}$$

In the multi-hop scenario, Eq. 24 replaces Eq. 11. Empirically, we found the amount of two-hop knowledge is substantially larger than that of one-hop knowledge, and hence introduces noise and additional computational complexity. To address these issues, we choose the top 100 two-hop knowledge pieces that are most similar to the dialogue context based on sentence-transformer scores for our experiments.

## 4 Experimental Setup

### 4.1 Datasets and Baselines

**Datasets.** Experiments are conducted on open-domain conversations extracted from Reddit (Zhou et al., 2018). ConceptNet (Speer et al., 2016) is used as the commonsense knowledge base, which consists of $120, 850$ triples, $21, 471$ entities, and $44$ relations. The knowledge base contains not only world facts, but also common concepts. Each single-round conversation pair is preserved if it can be connected by at least one knowledge triple. The dataset has $3, 384, 185/10, 000/20, 000$ conversations for training/evaluation/testing, respectively.

**Baselines.** We compare our model against five competitive baselines used in this task. There are some similar works, e.g. (Yu et al., 2022; Wu et al., 2022), which use external resources of documents or other kind of knowledge other than graph knowledge. They cannot be considered as our baseline models. Our research focuses on exploring a more efficacious approach for the integration of heterogeneous features within a language model framework. Consequently, large-scale language models, exemplified by ChatGPT [2], are neither employed as the primary language model in our experiments nor included within the baseline models under examination.

- **Seq2seq** (Sutskever et al., 2014): A widely used encoder-decoder in conversational systems.
- **MemNet** (Ghazvininejad et al., 2018): A model which uses MemNet to store knowledge triples.
- **CopyNet** (Zhu et al., 2017): A model which copies concepts in knowledge triples to generate responses.
- **CCM** (Zhou et al., 2018): The SOTA model for one-hop knowledge-enhanced dialogue which leverages two graph-attention mechanisms and

---

[2]ChatGPT, a recent language model release by https://chat.openai.com/, boasts a parameter count approximately 100 times greater than that of our base language model, UniLM.

CopyNet to model one-hop knowledge triples and incorporate knowledge concepts into responses.

- **ConceptFlow** (Zhang et al., 2020): The SOTA model for multi-hop knowledge-enhanced dialogue which has a similar method to CCM but uses additional graph attention to model two-hop knowledge triples.

## 4.2 Training Details and Parameters

UniLM-base-cased is used as the pre-trained language model. It has 12 BERT-block layers featuring 12 attention heads in each layer. The first six layers of the model are considered to be an encoder and the last six layers a decoder. The word embedding size is 768. The conversations and knowledge triples share the same BERT embedding layer, with a maximum length of 512. The hidden representation of the sixth layer is used to facilitate the 2-layer knowledge aggregation model. An Adam optimizer is used with a batch size of 36. The learning rate is $5e^{-5}$. The model was trained on a *Tesla V100* machine for approximately 7 days, and 20 epochs.

## 4.3 Evaluation Protocol

**Automatic Evaluation Metrics.** We follow (Zhou et al., 2018) and (Galley et al., 2018) in adopting the metrics of perplexity (PPL) (Serban et al., 2016) and Entity Score (ES), and follow (Galley et al., 2018) in adopting BLEU (Papineni et al., 2002), NIST (Doddington, 2002), METEOR (Lavie and Agarwal, 2007), Dist, and Ent (Zhang et al., 2018), where the Entity Score measures the average number of entities per response and others measure the quality of generated responses. BLEU, NIST, and METEOR are calculated between generated responses and golden responses, whilst Dist and Ent are calculated within generated responses.

**Human Evaluation.** Pair-wise comparisons are conducted with the most competitive baseline and the ablation model by five evaluators giving their preference of response on 100 randomly collected samples, regarding two aspects: the *appropriateness* (whether the response is appropriate in the context) and *informativeness* (whether the response contains new information).

## 5 Experimental Results

### 5.1 Automatic Evaluation

**Referenced Metrics.** The experimental results shown in Table 1 comprehensively measure the

quality of the generated responses. It can be observed that our CADGE model (which uses one-hop knowledge) outperforms most of the baselines. For instance, it outperforms CCM, one of the SOTA models using one-hop knowledge, on all metrics, obtaining at least twice the scores of the CCM (for BLEU-4, the difference is even almost four times). When compared to ConceptFlow, a SOTA model that exploits multi-hop knowledge, CADGE is still able to perform better (on over half of the metrics) or give comparable performance.

Given that the baselines contain the most representative framework for encoding heterogeneous features with separate encoders (i.e. CCM), the results clearly show the effectiveness of our knowledge aggregation mechanism, which better captures the heterogeneous features from the posts and knowledge facts with unified feature encoding and knowledge aggregation, and hence improves the quality of the generated responses. The ablation experiments further demonstrate the advances of the knowledge aggregation mechanism. Our context-aware GAT largely contributes to the improvement in performance, which can be observed in the comparison with - *w/o ca-gat*. Additionally, we also tried to allow neural networks to understand the semantics by directly coagulating the features of flattened triples - *w/o aggregation*, where the performance drops significantly, indicating the layer forward aggregation process is a key factor to the understanding of semantics contained in graph knowledge. By incorporating the enhanced two-hop knowledge, CADGE achieves universal performance gains on all metrics, further demonstrating the usefulness of incorporating multi-hot knowledge.

**Unreferenced Metrics.** We also examine the quality of the generated responses with unreferenced metrics that measure diversity and informativeness (entity score). As the results show in Table 2, both language diversity and informativeness are substantially improved with our proposed knowledge aggregation framework. For example, the diversity score of our model is on par with that of the SOTA model (ConceptFlow). When two-hop knowledge is incorporated, the scores of CADGE are almost double that of ConceptFlow, which also uses multi-hop knowledge.

These strong results demonstrate our model offers a substantial improvement over existing approaches when considering the language quality and relevance of generated responses, and matches

| Model | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | NIST-1 | NIST-2 | NIST-3 | NIST-4 | METEOR |
|---|---|---|---|---|---|---|---|---|---|
| Seq2Seq | 0.1702 | 0.0579 | 0.0226 | 0.0098 | 1.0230 | 1.0963 | 1.1056 | 1.1069 | 0.0611 |
| MemNet | 0.1741 | 0.0604 | 0.0246 | 0.0112 | 1.0975 | 1.1847 | 1.1960 | 1.1977 | 0.0632 |
| CopyNet | 0.1589 | 0.0549 | 0.0226 | 0.0106 | 0.9899 | 1.0664 | 1.0770 | 1.0788 | 0.0610 |
| CCM | 0.1413 | 0.0484 | 0.0192 | 0.0084 | 0.8362 | 0.9000 | 0.9082 | 0.9095 | 0.0630 |
| ConceptFlow | **0.2451** | **0.1047** | 0.0493 | 0.0246 | <u>1.6137</u> | 1.7956 | 1.8265 | 1.8329 | <u>0.0942</u> |
| CADGE | 0.2078 | 0.0967 | <u>0.0551</u> | <u>0.0326</u> | 1.5566 | <u>1.8113</u> | <u>1.8609</u> | <u>1.8683</u> | 0.0893 |
| - w/o es-loss | 0.2024 | 0.0937 | 0.0525 | 0.0315 | 1.5114 | 1.7421 | 1.7826 | 1.7878 | 0.0895 |
| - w/o aggregation | 0.1941 | 0.0920 | 0.0528 | 0.0322 | 1.4672 | 1.6994 | 1.7421 | 1.7477 | 0.0861 |
| - w/o ca-gat | 0.2019 | 0.0730 | 0.0305 | 0.0138 | 1.3562 | 1.4919 | 1.5082 | 1.5101 | 0.0796 |
| - w/ two hops | <u>0.2197</u> | <u>0.1011</u> | **0.0558** | **0.0328** | **1.6689** | **1.9171** | **1.9606** | **1.9661** | **0.1053** |

Table 1: Automatic evaluation on popular reference-based metrics used in the task of open domain dialogue. The best performing model is highlighted in **bold**, and the second best is <u>underlined</u>. **- w/o es-loss** denotes the ablated model without the auxiliary entity selection task; **- w/o aggregation** denotes the model without the feature aggregation process (which is implemented by directly mean pooling the features of flattened triples without our two layer forward aggregation process); **- w/o ca-gat** denotes the model without our proposed context-aware GAT introduced in subsection 3.2; **- w/ two hops** denotes the model expanded by two-hop knowledge introduced in subsection 3.5.

| Model | Dist-1 | Dist-2 | Ent-4 |
|---|---|---|---|
| Seq2Seq | 0.0123 | 0.0525 | 7.665 |
| MemNet | 0.0211 | 0.0931 | 8.418 |
| CopyNet | 0.0223 | 0.0988 | 8.422 |
| CCM | 0.0146 | 0.0643 | 7.847 |
| Conceptflow | 0.0223 | 0.1228 | <u>10.270</u> |
| CADGE | 0.0288 | 0.1136 | 10.141 |
| - w/d es-loss | 0.0326 | <u>0.1242</u> | 9.445 |
| - w/o aggregation | <u>0.0340</u> | 0.1234 | 8.968 |
| - w/d ca-gat | 0.0189 | 0.0755 | 9.599 |
| - w/ two hops | **0.0461** | **0.2702** | **11.626** |

Table 2: Automatic evaluation on unreferenced metrics.



Figure 4: The learned attention probability density curves on knowledge facts.

better with the golden reference responses. When generating responses only with the UniLM model, performance on all metrics drops substantially, further demonstrating that the proposed Context-aware GAT contributes immensely to generating informative and high-quality responses via effective aggregation of knowledge triples. Both the referenced and unreferenced metrics indicate that with the improvement in heterogeneous feature capturing and global feature aggregation, CADGE can better exploit background knowledge to generate more high-quality and human-like responses.

## 5.2 Analysis of the Knowledge Aggregation Mechanism

**Perplexity and Entity Score.** Based on the frequency of words in the posts, we divide the test set into four sections (high, middle, low, and OOV) in order to evaluate the performance and robustness of each model when faced with frequently seen dialogues as well as uncommon dialogues. For a fair comparison, we limit the retrieved knowledge to
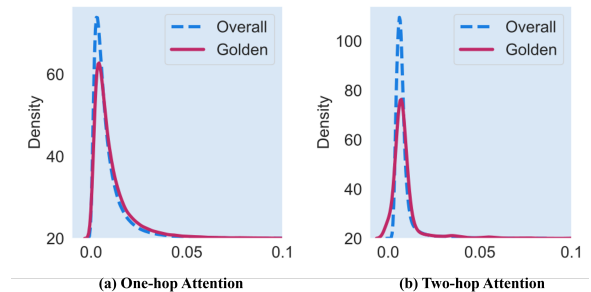
one-hop as not every baseline is able to incorporate multi-hop knowledge (e.g., CCM). As shown in Table 3, our model achieves the lowest perplexity and the highest entity scores for all frequency groups. The lowest perplexity indicates that the proposed model achieves the best predictive performance of the language model and generates a more fluent response than other baselines, while the best entity scores indicate that the proposed model better exploits graph features to select appropriate entities contained in the post. For the ablation study, we compare CADGE to the base model UniLM,[3] which is a pre-trained language model without the Context-aware GAT. The substantial performance gain of CADGE over UniLM demonstrates the importance of leveraging global features obtained by graph knowledge to improve both the model's understanding and generation ability.

**Attention Distribution on Knowledge.** In order to test whether our model has learned to place more at-

---

[3]The ablated model **- w/o ca-gat** is regarded as the base model UniLM, which works without graph knowledge.

| Model | Overall | | High Freq. | | Medium Freq. | | Low Freq. | | OOV | |
|---|---|---|---|---|---|---|---|---|---|---|
| | PPL↓ | ES↑ | PPL↓ | ES↑ | PPL↓ | ES↑ | PPL↓ | ES↑ | PPL↓ | ES↑ |
| Seq2Se | 47.02 | 0.72 | 42.41 | 0.71 | 47.25 | 0.74 | 48.61 | 0.72 | 49.96 | 0.67 |
| MemNet | 46.85 | 0.76 | 41.93 | 0.76 | 47.32 | 0.79 | 48.86 | 0.76 | 49.52 | 0.71 |
| CopyNet | 40.27 | 0.96 | 36.26 | 0.91 | 40.99 | 0.97 | 42.09 | 0.96 | 42.24 | 0.96 |
| CCM | 39.18 | 1.18 | 35.36 | 1.16 | 39.64 | 1.19 | 40.67 | 1.20 | 40.87 | 1.16 |
| CADGE | **33.99** | **1.39** | **31.50** | **1.49** | **34.39** | **1.43** | **34.67** | **1.35** | **35.56** | **1.29** |
| - w/o es-loss | 34.73 | 1.28 | 32.31 | 1.36 | 35.18 | 1.33 | 35.41 | 1.24 | 35.19 | 1.19 |
| - w/o aggregation | 34.71 | 1.35 | 32.25 | 1.42 | 35.16 | 1.39 | 35.36 | 1.31 | 35.62 | 1.27 |
| - w/o ca-gat | 36.51 | 1.03 | 33.82 | 1.10 | 37.02 | 1.06 | 37.23 | 1.01 | 38.12 | 0.95 |

Table 3: Automatic evaluation on the metrics of *perplexity* (↓) and *entity score* (↑). The experiment is set up with one-hop knowledge. Therefore ConceptFlow, which needs two-hop knowledge, is excluded in this experiment. The test set (**Overall**) is categorised into 4 sub-datasets with different frequencies (**Freq.** and **OOV** (out of vocabulary)) of the entities included in the posts. The overall PPL and ES of **ConceptFlow** are 36.51 and 1.03, respectively. The overall PPL and ES of Cadge **- w/ two hops** are 29.90 and 1.68, respectively. Since ConceptFlow did not evaluate frequency grouped test data on two-hop data, we only compare models with one-hop data here.
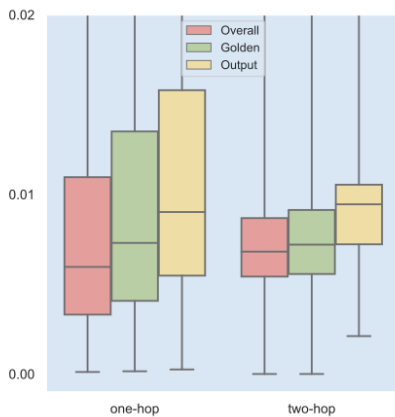


Figure 5: A box plot to analyse attention scores learned by context-aware GAT to aggregate features from one-hop and two-hop knowledge. **Overall**: average attention of all knowledge; **Golden**: average attention of golden knowledge; **Output**: average attention of knowledge in generated responses.

tention on golden knowledge facts for dialogue generation, we draw probability density curves to compare the attention distribution of golden knowledge (i.e. retrieved knowledge facts that appear in *reference responses*) and overall knowledge (knowledge facts retrieved from posts). Figure 4 illustrates the result with one-hop knowledge aggregation, and Figure 4 with two-hop. It can be observed that Context-aware GAT is able to learn to select more related knowledge facts for dialogue generation, as demonstrated by the curves showing that golden knowledge facts have a higher probability of having higher attention scores. In other words, our graph model is able to obtain an aggregated representation that places more focus on relevant knowledge for response generation.

**Statistics for Attention Scores.** To better analyse the statistics of the learned attention scores during the knowledge aggregation in our model, we further draw a box plot to compare the attention scores of different knowledge facts, with the results shown in Figure 5. According to the attention scores distribution, the knowledge facts in the output have higher attention than other retrieved knowledge, meaning the model has more confidence to select related knowledge to generate responses.[4] With respect to the attention on the golden knowledge facts, they are substantially different from other retrieved knowledge, which demonstrates that with the knowledge aggregation process, our framework learned the correct features to represent knowledge facts, leading to more appropriate selections over retrieved knowledge facts.

## 5.3 Human Evaluation

We also conducted human evaluation to further consolidate our model performance. The results are presented in Table 4, which, in accordance with the previously presented automatic metrics, demonstrates that our model outperforms the SOTA baselines on both *appropriateness* and *informativeness*, and proves the effectiveness of the proposed Context-Aware GAT. Under the condition of either one-hop knowledge or two-hop knowledge, CADGE achieves significant improvements in producing more informative and appropriate responses, owing to the proposed context-aware knowledge aggregation framework.

---

[4]If the generated knowledge facts have the same distribution as the overall, this means that the model is confused when selecting relevant knowledge facts.

| Choice % | CADGE$_{one\_hop}$ vs CCM | | | CADGE$_{one\_hop}$ vs - w/o ca-gat | | | CADGE$_{two\_hops}$ vs ConceptFlow | | |
|---|---|---|---|---|---|---|---|---|---|
| | CADGE$_{one\_hop}$ | CCM | Kappa | CADGE$_{one\_hop}$ | - w/o ca-gat | Kappa | CADGE$_{two\_hops}$ | Conceptflow | Kappa |
| App. | 66.0 | 34.0 | 0.367 | 58.1 | 41.9 | 0.323 | 64.7 | 35.3 | 0.321 |
| Inf. | 63.3 | 36.7 | 0.278 | 60.1 | 39.9 | 0.318 | 64.9 | 35.1 | 0.304 |

Table 4: Human Evaluation w.r.t. *appropriateness* and *informativeness*. The score is the percentage that the proposed model wins against its competitor. *Kappa* denotes Fleiss' Kappa (Fleiss, 1971), which indicates all of our evaluation annotations reach a fair agreement. The proposed model is significantly better (sign test, $p < 0.005$).
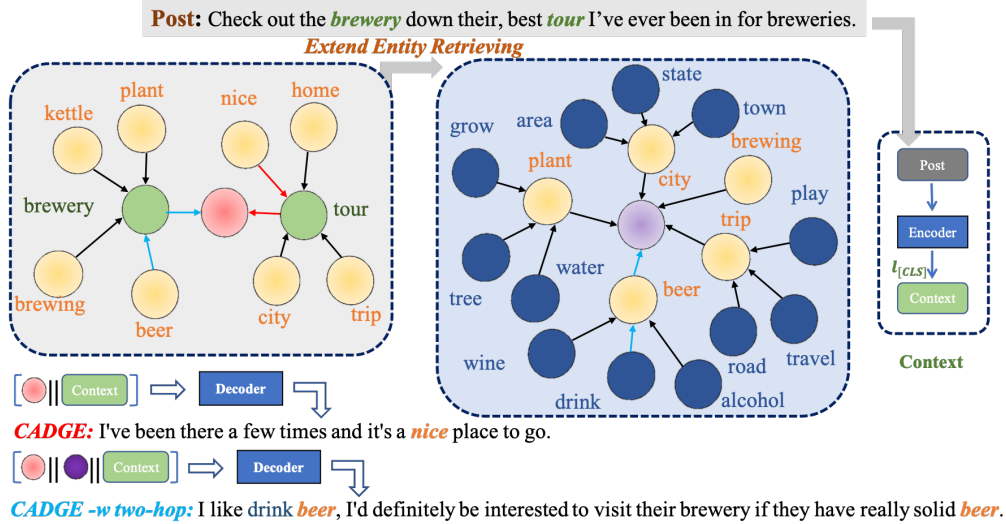


Figure 6: Visualization of the knowledge aggregation process with an example.

## 5.4 Knowledge Aggregation Process.

In Figure 6, we illustrate an example of the knowledge aggregation process of our framework, where the left subgraph represents the one-hop knowledge aggregation (i.e. yellow nodes) and the right subgraph represents the additional knowledge aggregation attending to the second-hop knowledge (i.e. blue nodes). As mentioned in §3, CADGE aggregates features layer by layer. For one-hop CADGE, the aggregated representation (the red node) of all retrieved knowledge facts is concatenated with the context features of the post, and fed into the neural decoder to generate responses. When incorporating two-hop knowledge, CADGE exploits a similar mechanism, and we obtain an additional knowledge representation (the purple node) for response generation. It can be seen from the example that when CADGE only uses one-hop knowledge, it selects "nice" from the graph which is subsequently used to generate a response. When two-hop knowledge is available, CADGE selects "beer" from the one-hop graph and "drink" from the two-hop graph, improving informativeness and making the response more interesting. We also provide a detailed qualitative analysis of sample responses from the one-hop and two-hop knowledge

experiments in Appendix B.

## 6 Conclusion

In this paper, we proposed a novel knowledge aggregation framework for the knowledge graph enhanced dialogue generation task. This framework implements a Context-aware GAT which applies representation learning of the heterogeneous features from graph knowledge text, and the neural networks effectively learn to incorporate globally aggregated features to enhance response generation with rich representations. Extensive experiments are conducted to demonstrate that our framework outperforms SOTA baselines on both automatic and human evaluation, as the proposed Context-Aware GAT largely improved the semantic understanding of both graph and text knowledge to enhance the appropriateness and informativeness of generated responses. The expansion of Context-Aware GAT to two-hop knowledge also indicates the robustness and effectiveness of our framework in increasing the amount of grounded graph knowledge in responses. We hope that our proposed framework can benefit research in all text generation tasks where knowledge graphs are incorporated, and transferable research will be continued in further work.

## Acknowledgements

## Ethics Statement

All work presented within this paper is in line with the ethical code of conduct of both ACL and the institutions of the authors. In this work we present a method to increase the level of world knowledge in dialogue system responses. In turn, this results in the applications being more useful to the end user, and being better positioned to answer a range of topics. However, we acknowledge that using similar approaches to incorporate knowledge into dialogue systems should be cautious of the veracity and validity of the utilised "knowledge" in order to avoid issues relating to misinformation. We do not motivate this work in terms of a specific application, and instead present a method for incorporating knowledge graph structured information in general.

## References

Noor Fazilla Abd Yusof, Chenghua Lin, and Frank Guerin. 2017. Analysing the causes of depressed mood from depression vulnerable individuals. In *Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017)*, pages 9–17.

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot.

Siqi Bao, Huang He, Fan Wang, and Hua Wu. 2019. PLATO: pre-trained dialogue generation model with discrete latent variable. *CoRR*, abs/1910.07931.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2020. Plato-2: Towards building an open-domain chatbot via curriculum learning. *arXiv preprint arXiv:2006.16779*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, page 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *Advances in Neural Information Processing Systems*, 32.

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

Michel Galley, Chris Brockett, Xiang Gao, Bill Dolan, and Jianfeng Gao. 2018. End-to-end conversation modeling: Moving beyond chitchat.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Henglin Huang, Chen Tang, Tyler Loakman, Frank Guerin, and Chenghua Lin. 2022. Improving Chinese story generation via awareness of syntactic dependencies and semantics. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 178–185, Online only. Association for Computational Linguistics.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. *CoRR*, abs/1902.04911.

Weizhe Lin, B-H Tseng, and Bill Byrne. 2021. Knowledge-aware graph-enhanced gpt-2 for dialogue state tracking. *ArXiv*, abs/2104.04466.

Ehsan Lotfi, Maxime De Bruyn, Jeska Buhmann, and Walter Daelemans. 2021. Teach me what to say and I will learn what to pick: Unsupervised knowledge selection through response generation with pretrained

generative models. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 254–262, Online. Association for Computational Linguistics.

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

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 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, Online. Association for Computational Linguistics.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2009. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80.

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.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. *ArXiv*, abs/1612.03975.

Manisha Srivastava, Yichao Lu, Riley Peschon, and Chenyang Li. 2021. Pretrain-finetune based training of task-oriented dialogue systems in a real-world setting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 34–40.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Chen Tang, Chenghua Lin, Henglin Huang, Frank Guerin, and Zhihao Zhang. 2022a. EtriCA: Event-triggered context-aware story generation augmented by cross attention. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5504–5518, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chen Tang, Tyler Loakman, and Chenghua Lin. 2024. A cross-attention augmented model for event-triggered context-aware story generation. *Computer Speech & Language*, 88:101662.

Chen Tang, Shun Wang, Tomas Goldsack, and Chenghua Lin. 2023a. Improving biomedical abstractive summarisation with knowledge aggregation from citation papers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 606–618, Singapore. Association for Computational Linguistics.

Chen Tang, Hongbo Zhang, Tyler Loakman, Chenghua Lin, and Frank Guerin. 2023b. Enhancing dialogue generation via dynamic graph knowledge aggregation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4604–4616, Toronto, Canada. Association for Computational Linguistics.

Chen Tang, Hongbo Zhang, Tyler Loakman, Chenghua Lin, and Frank Guerin. 2023c. Terminology-aware medical dialogue generation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Chen Tang, Zhihao Zhang, Tyler Loakman, Chenghua Lin, and Frank Guerin. 2022b. NGEP: A graph-based event planning framework for story generation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 186–193, Online only. Association for Computational Linguistics.

Yi-Lin Tuan, Yun-Nung Chen, and Hung-yi Lee. 2019. Dykgchat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. *arXiv preprint arXiv:1910.00610*.

Dingmin Wang, Chenghua Lin, Qi Liu, and Kam-Fai Wong. 2021. Fast and scalable dialogue state tracking with explicit modular decomposition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 289–295.

Sixing Wu, Minghui Wang, Ying Li, Dawei Zhang, and Zhonghai Wu. 2022. Improving the applicability of knowledge-enhanced dialogue generation systems by using heterogeneous knowledge from multiple sources. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1149–1157.

Yu Xie, Bin Yu, Shengze Lv, Chen Zhang, Guodong Wang, and Maoguo Gong. 2021. A survey on heterogeneous network representation learning. *Pattern Recognition*, 116:107936.

Bohao Yang, Chen Tang, and Chenghua Lin. 2024a. Improving medical dialogue generation with abstract meaning representations. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11826–11830. IEEE.

Bohao Yang, Chen Tang, Kun Zhao, Chenghao Xiao, and Chenghua Lin. 2024b. Effective distillation of

table-based reasoning ability from LLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5538–5550, Torino, Italia. ELRA and ICCL.

Shiquan Yang, Rui Zhang, and Sarah Monazam Erfani. 2020. Graphdialog: Integrating graph knowledge into end-to-end task-oriented dialogue systems. *ArXiv*, abs/2010.01447.

Wenhao Yu, Chenguang Zhu, Lianhui Qin, Zhihan Zhang, Tong Zhao, and Meng Jiang. 2022. Diversifying content generation for commonsense reasoning with mixture of knowledge graph experts. In *Proceedings of the 2nd Workshop on Deep Learning on Graphs for Natural Language Processing (DLG4NLP 2022)*, pages 1–11, Seattle, Washington. Association for Computational Linguistics.

Chengkun Zeng, Guanyi Chen, Chenghua Lin, Ruizhe Li, and Zhi Chen. 2021. Affective decoding for empathetic response generation. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 331–340.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2031–2043, Online. Association for Computational Linguistics.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. *Advances in Neural Information Processing Systems*, 31.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation.

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.

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

## A Related Work

**Knowledge Enhanced Dialogue Generation.** Recently, much work has focused on augmenting dialogue systems with additional background knowledge. Such works can be divided into dialogue systems augmented with unstructured knowledge, and

| #1 Post | So if op has ever taken a road trip he was at more of a risk than the surgery he's gonna have :d |
|---|---|
| CADGE$_{one\_hop}$ | I'm not saying it's safe, but it's not like he's going to be in a car accident. |
| CCM | I'm not sure if he 's a surgeon or not . |
| - *w/o ca-gat* | I'm not saying he's not going to be taking a road trip. |
| **#2 Post** | **Because i haven't heard it on the radio enough.** |
| CADGE$_{one\_hop}$ | I've heard it on radio. it's a pretty good song |
| CCM | I'm not sure if it's a radio station or not. i 'm not sure if it 's a radio station or not. |
| - *w/o ca-gat* | I've heard it on the radio. |
| **#3 Post** | **We will have to agree or disagree mate!** |
| CADGE$_{one\_hop}$ | I'm not disagreeing, i'm just saying that it's not a valid argument. |
| CCM | I'm not saying it 's a bad argument , but i don't think it 's a good idea to be a fan of the game. |
| - *w/o ca-gat* | I'm not your mate, pal. |
| **#4 Post** | **What do you mean by the auto dude?** |
| CADGE$_{one\_hop}$ | I mean the guy who's in the car. |
| CCM | I'm not sure . i 'm not sure if i 'm wrong . i 'm just saying that the guy is a guy who doesn't know what he 's talking about . |
| - *w/o ca-gat* | Auto man. |

Table 5: Sample responses from the one-hop knowledge experiments, comparing CADGE with the SOTA CCM, and its ablation model UniLM. Tokens in red are entities in knowledge triples.

those augmented with structured knowledge. With unstructured knowledge, (Komeili et al., 2022) models web page information and feeds it into a language model. (Ghazvininejad et al., 2018) and (Lotfi et al., 2021) encode the filtered factual statements with a specific encoder and then pass them into the decoder along with context. (Lian et al., 2019) use context to aggregate knowledge statements and find that aggregated knowledge gives better results than filtered knowledge. Regarding structured knowledge, graph neural networks (Scarselli et al., 2009) are usually used to embed graph information to input into a language model. (Zhou et al., 2018) uses GRUs and two graph attention modules to select appropriate triples to incorporate into responses. In order to exploit the benefits of multi-hop knowledge, (Zhang et al., 2020) adds an attention mechanism in a similar way to filter the appropriate knowledge. Finally, (Tuan et al., 2019) proposes a model which selects the output from a sequence-to-sequence model and a multi-hop reasoning model at each time step.

## B Case Study

Several post-response pairs are presented to compare generated responses by the proposed CADGE, the ablated model UniLM, and the SOTA models of CCM and ConceptFlow, in different external

| #1 Post | My brain must not like music, i've never gotten chills from a song. |
|---|---|
| CADGE$_{two\_hop}$ | I don't really listen to music at all but i have found some songs that you may enjoy like. |
| ConceptFlow | The whole album is a great song. |
| CADGE$_{one\_hop}$ | I don't even listen to music, I don't love music and i have no idea how to get chills from something . |
| #2 Post | The rest of the series will be released for free as well. |
| CADGE$_{two\_hop}$ | I don't know about the rest of the collection, i only really read the second book . |
| ConceptFlow | The whole album is a great song. |
| CADGE$_{one\_hop}$ | I don't really care if it's on sale, i can just play it on steam. |
| #3 Post | Check out the brewery down their, best tour ive ever been in for breweries |
| CADGE$_{two\_hop}$ | I like drink beer, i'd definitely be interested to visit their brewery if they have really solid beer. |
| ConceptFlow | I've been to the brewery. it's a blast to see. |
| CADGE$_{one\_hop}$ | i've been there a few times and it's a nice place to go. |
| #4 Post | I'm thoroughly disappointed that i don't have work - appropriate white clothing for today. |
| CADGE$_{two\_hop}$ | I just go through some tweet to find a better outfit to wear. |
| ConceptFlow | I'll have to check out my new white shirts for the next day |
| CADGE$_{one\_hop}$ | You have to wear a white shirt to work for a few days a week. |

Table 6: Sample responses from the two-hop knowledge experiments, comparing CADGE$_{two\_hop}$ with the SOTA model ConceptFlow, and its one-hop ablation model CADGE$_{one\_hop}$. Tokens in red are entities belonging to the one hop knowledge, while tokens in blue belong to the second hop.

knowledge settings (one-hop or two-hop). Table 5 presents comparisons based on one-hop knowledge. Considering the ablation of external knowledge, it can be observed that without the representations for the knowledge facts, the UniLM model struggled to understand the context semantics and provide informative responses. For example, in the first post, the one-hop CADGE understands that the focus of the post is on "risk", while UniLM considers it to be on "road trip". In the third post, the one-hop CADGE understands that the focus of the post is on "agree", while UniLM considers it to be "mate".

When we consider the effectiveness of knowledge fact exploitation, the difference can be observed in generated responses between the one-hop CADGE and the CCM model. Responses from CADGE appear to be more logical and fluent than CCM. For instance, in the fourth post, the one-hop CADGE understands the phrase "auto dude" and gives an accurate explanation, instead of saying "not sure" as CCM does. The same phenomenon also appears in the first and second posts, which demonstrates that with the proposed knowledge aggregation framework, CADGE is more able to understand knowledge facts, and provide more informative and appropriate answers with this knowledge.

In regards to the expansion on two-hop knowledge, our context GAT sustains the effectiveness and efficiency of knowledge representation learning. The additional comparisons are compared among CADGE$_{one\_hop}$, CADGE$_{two\_hop}$, and ConceptFlow in Table 6. It can be observed that when the knowledge amount increases, CADGE$_{two\_hop}$ is better able to consider background knowledge when generating responses. For example, in the second and third post, CADGE$_{two\_hop}$ considers more retrieved knowledge facts to generate a response which results in responses with better quality, and that are more informative. In addition, the extra knowledge also gives more context semantics leading to better understanding of the dialogues. For instance, in all of the aforementioned cases, compared to one-hop CADGE and ConceptFlow, the two-hop CADGE chooses more informative concepts from all available knowledge, making the generated responses more interesting.