DyKgChat: Benchmarking Dialogue Generation Grounding on Dynamic Knowledge Graphs

Yi-Lin Tuan  Yun-Nung Chen  Hung-yi Lee
National Taiwan University, Taipei, Taiwan
pascaltuan@gmail.com  y.v.chen@ieee.org  hungyilee@ntu.edu.tw

Abstract

Data-driven, knowledge-grounded neural conversation models are capable of generating more informative responses. However, these models have not yet demonstrated that they can zero-shot adapt to updated, unseen knowledge graphs. This paper proposes a new task about how to apply dynamic knowledge graphs in neural conversation model and presents a novel TV series conversation corpus (DyKgChat) for the task. Our new task and corpus aids in understanding the influence of dynamic knowledge graphs on responses generation. Also, we propose a preliminary model that selects an output from two networks at each time step: a sequence-to-sequence model (Seq2Seq) and a multi-hop reasoning model, in order to support dynamic knowledge graphs. To benchmark this new task and evaluate the capability of adaptation, we introduce several evaluation metrics and the experiments show that our proposed approach outperforms previous knowledge-grounded conversation models. The proposed corpus and model can motivate the future research directions.

1 Introduction

In the chit-chat dialogue generation, neural conversation models (Sutskever et al., 2014; Sordoni et al., 2015; Vinyals and Le, 2015) have emerged for its capability to be fully data-driven and end-to-end trained. While the generated responses are often reasonable but general (without useful information), recent work proposed knowledge-grounded models (Eric et al., 2017; Ghazvininejad et al., 2018; Zhou et al., 2018b; Qian et al., 2018) to incorporate external facts in an end-to-end fashion without hand-crafted slot filling. Effectively combining text and external knowledge graphs have also been a crucial topic in question answering (Yin et al., 2016; Hao et al., 2017; Levy et al., 2017; Sun et al., 2018; Das et al., 2019). Nonetheless, prior work rarely analyzed the model capability of zero-shot adaptation to dynamic knowledge graphs, where the states/entities and their relations are temporal and evolve as a single time scale process. For example, as shown in Figure 1, the entity Jin-Xi was originally related to the entity Feng, Ruozhao with the type EnemyOf, but then evolved to be related to the entity Nian, Shilan.

The goal of this paper is to facilitate knowledge-grounded neural conversation models to learn and zero-shot adapt with dynamic knowledge graphs. To our observation, however, there is no existing conversational data paired with dynamic knowledge graphs. Therefore, we collect a TV series corpus—DyKgChat, with facts of the fictitious life of characters. DyKgChat includes a Chinese palace drama Hou Gong Zhen Huan Zhuang (HGZH), and an English sitcom Friends, which contain dialogues, speakers, scenes (e.g., the places and listeners), and the corresponded knowl-

Figure 1: An example of an ideal conversation model with dynamic knowledge graphs.

1The data and code are available in https://github.com/Pascalson/DyKGChat.
edge graphs including explicit information such as the relations \textit{FriendOf}, \textit{EnemyOf}, and \textit{ResidenceOf} as well as the linked entities. Table 1 shows some examples from DyKgChat.

| HGZH | Zhen-Huan: She must be frightened. It should blame me. I should not ask her to play chess. |
|------|-----------------------------------------------------------------------------------|
|      | Joey: Alright, Joey, be nice. So does he have a hump? A hump and a hairpiece?      |

Table 1: Examples of DyKgChat corpus.

- A new task, \textit{dynamic knowledge-grounded conversation generation}, is proposed.
- A newly-collected TV series conversation corpus \textit{DyKgChat} is presented for the target task.
- We benchmark the task by comparing many prior models and the proposed quick adaptive dynamic knowledge-grounded neural conversation model (Qadpt), providing the potential of benefiting the future research direction.

2 Task Description

For each single-turn conversation, the input message and response are respectively denoted as \( x = \{x_t\}_{t=1}^m \) and \( y = \{y_t\}_{t=1}^n \), where \( m \) and \( n \) are their lengths. Each turn \((x, y)\) is paired with a knowledge graph \( K \), which is composed of a collection of triplets \((h, r, t)\), where \( h, t \in V \) (the set of entities) and \( r \in L \) (the set of relationships). Each word \( y_t \) in a response belongs to either generic words \( \mathcal{W} \) or knowledge graph entities \( V \). The task is two-fold:

1. Given an input message \( x \) and a knowledge graph \( K \), the goal is to generate a sequence \( \{\hat{y}_t\}_{t=1}^n \) that is not only as similar as possible to the ground-truth \( \{y_t\}_{t=1}^n \), but contains correct knowledge graph entities to reflect the information.
2. After a knowledge graph is updated to \( K' \), where some triplets are revised to \((h, r, t')\) or \((h, r', t)\), the generated sequence should contain correspondent knowledge graph entities in \( K' \) to reflect the updated information.

2.1 Evaluation Metrics

To evaluate dynamic knowledge-grounded conversation models, we propose two types of evaluation metrics for validating two aspects described above.

2.1.1 Knowledge Entity Modeling

There are three metrics focusing on the knowledge-related capability.
Knowledge word accuracy (KW-Acc). Given the ground-truth sentence as the decoder inputs, at each time step, it evaluates how many knowledge graph entities are correctly predicted.

\[
\text{KW-Acc} = \sum_{t=1}^{n} P(\hat{y}_t = y_t | y_1 y_2 \ldots y_{t-1}, y_t \in \mathcal{V}).
\]

For example, after perceiving the partial ground-truth response “If Jin-Xi not in” and knowing the next word should be a knowledge graph entity, KW-Acc measures if the model can predict the correct word “Yongshou Palace”.

Knowledge and generic word classification (KW/Generic). Given the ground-truth sentence as the decoder inputs, at each time step, it measures the capability of predicting the correct class (a knowledge graph entity or a generic word) and adopts micro-averaging. The true positive, false negative and false positive are formulated as:

\[
\begin{align*}
\text{TP} &= |\{ t \mid \hat{y}_t \in \mathcal{V}, y_t \in \mathcal{V}\}|, \\
\text{FN} &= |\{ t \mid \hat{y}_t \in \mathcal{W}, y_t \in \mathcal{V}\}|, \\
\text{FP} &= |\{ t \mid \hat{y}_t \in \mathcal{V}, y_t \in \mathcal{W}\}|,
\end{align*}
\]
\[
\hat{y}_t \sim P(\cdot | y_1 y_2 \ldots y_{t-1}).
\]

Generated knowledge words (Generated-KW). Considering the knowledge graph entities in the reference \( y = \{ y_t \}_{t=1}^{n} \) as positives, in the inference stage, we use the generated knowledge entities to compute true positive, false positive, and true negative, and adopt micro-averaging.

\[
\begin{align*}
\text{TP} &= |\{ \hat{y}_t \in \{ y_t \}_{t=1}^{n}, \hat{y}_t \in \mathcal{V}\}_{t=1}^{n}|, \\
\text{FN} &= |\{ y_t \notin \{ \hat{y}_t \}_{t=1}^{n}, y_t \in \mathcal{V}\}_{t=1}^{n}|, \\
\text{FP} &= |\{ \hat{y}_t \notin \{ y_t \}_{t=1}^{n}, \hat{y}_t \in \mathcal{V}\}_{t=1}^{n}|,
\end{align*}
\]
\[
\hat{y}_t \sim P(\cdot | y_1 y_2 \ldots y_{t-1}).
\]

For example, after input a sentence “Where’s JinXi?”, if a model generates “Hi, Zhen-Huan, JinXi is in Yangxin-Palace.” when reference is “JinXi is in Yongshou-Palace.”, where bolded words are knowledge entities. Recall is \( \frac{1}{2} \) and precision is \( \frac{1}{3} \).

2.1.2 Adaptation of Changed Knowledge Graphs

Each knowledge graph is randomly changed by (1) shuffling a batch (All), (2) replacing the predicted entities (Last1), or (3) replacing the last two steps of paths predicting the generated entities (Last2). We have two metrics focusing on the capability of adaptation.

Change rate. It measures if the responses are different from the original ones (with original knowledge graphs). The higher rate indicates that the model is more sensitive to a changed knowledge graph. Therefore, the higher rate may not be better, because some changes are worse. The following metric is proposed to deal with the issue, but change rate is still reported.

Accurate change rate. This measures if the original predicted entities are replaced with the hypothesis set, where this ensures that the updated responses generate knowledge graph entities according to the updated knowledge graphs. (1) In All, the hypothesis set is the collection of all entities in the new knowledge graph. (2) In Last1 and Last2, the hypothesis set is the randomly-selected substitutes.

3 DyKgChat Corpus

This section introduces the collected DyKgChat corpus for the target knowledge-grounded conversation generation task.

3.1 Data Collection

To build a corpus where the knowledge graphs would naturally evolves, we collect TV series conversations, considering that TV series often contain complex relationship evolution, such as
friends, jobs, and residences. We choose TV series with different languages and longer episodes. We download the scripts of a Chinese palace drama “Hou Gong Zhen Huan Zhuang” (HGZHZ; with 76 episodes and hundreds of characters) from Baidu Tieba, and the scripts of an English sitcom “Friends” (with 236 episodes and six main characters)\(^2\). Their paired knowledge graphs are manually constructed according to their wikis written by fans\(^3\). Noted that the knowledge graph of HGZHZ is mainly built upon the top twenty-five appeared characters.

The datasets are split 5% as validation data and 10% as testing data, where the split is based on multi-turn dialogues and balanced on speakers. The boundaries of dialogues are annotated in the original scripts. The tokenization of HGZHZ considers Chinese characters and knowledge entities; the tokenization of Friends considers space-separated tokens and knowledge entities. The data statistics after preprocessing is detailed in Table 2. The relation types \(r \in \mathcal{L}\) of each knowledge graph and their percentages are listed in Table 3, and the knowledge graph entities are plotted as word clouds in Figure 2.

3.2 Subgraph Sampling

Due to the excessive labor of building dynamic knowledge graphs aligned with all episodes, we currently collect a fixed knowledge graph \(G\) containing all information that once exists for each TV series. To build the aligned dynamic knowledge graphs, we sample the top-five shortest paths on knowledge graphs from each source to each target, where the sources are knowledge entities in the input message and the scene \(\{x_t \in \mathcal{V}\}\), and the targets are knowledge entities in the ground-truth response \(\{y_t \in \mathcal{V}\}\). We manually check whether the selected number of shortest paths are able to cover most of the used relational paths. The dynamic knowledge graphs are built based on an ensemble of the following possible subgraphs:

- The sample for each single-turn dialogue.
- The sample for each multi-turn dialogue.
- The manually-annotated subgraph for each period.

While the first rule is adopted for simplicity, the preliminary models should at least work on this type of subgraphs. The subgraphs are defined as the dynamic knowledge graphs \(\{\mathcal{K}\}\), which are updated every single-turn dialogue.

3.3 Data Analysis

Data imbalance. As shown in Table 2, the turns with knowledge graph entities are about 58.9% and 15.93% of HGZHZ and Friends respectively. Apparently in Friends, the training data with knowledge graph entities are too small, so fine-
Figure 4: The framework of the proposed model. The node E here is the symbol for the emperor.

4 Qadpt: Quick Adaptative Dynamic Knowledge-Grounded Neural Conversation Model

To our best knowledge, no prior work focused on dynamic knowledge-grounded conversation; thus we propose Qadpt as the preliminary model. As illustrated in Figure 4, the model is composed of (1) a Seq2Seq model with a controller, which decides to predict knowledge graph entities \( k \in \mathcal{V} \) or generic words \( w \in \mathcal{W} \), and (2) a reasoning model, which retrieves the relational paths in the knowledge graph.

4.1 Sequence-to-Sequence Model

Qadpt is based on a Seq2Seq model (Sutskever et al., 2014; Vinyals and Le, 2015), where the encoder encodes an input message \( x \) into a vector \( e(x) \) as the initial state of the decoder. At each time step \( t \), the decoder produces a vector \( d_t \) conditioned on the ground-truth or predicted \( y_1 y_2 \ldots y_{t-1} \). Note that we use gated recurrent unit (GRU) (Cho et al., 2014) in our experiments.

\[
e(x) = \text{GRU}(x_1 x_2 \ldots x_m) \quad (1)
\]

\[
d_t = \text{GRU}(y_1 y_2 \ldots y_{t-1}, e(x)) \quad (2)
\]

Each predicted \( d_t \) is used for three parts: output projection, controller, and reasoning. For output projection, the predicted \( d_t \) is transformed into a distribution \( w_t \) over generic words \( \mathcal{W} \) by a projection layer.

4.2 Controller

To decide which vocabulary set (knowledge graph entities \( \mathcal{V} \) or generic words \( \mathcal{W} \)) to use, the vector \( d_t \) is transformed to a controller \( c_t \), which is a widely-used component (Eric et al., 2017; Zhu et al., 2017; Ke et al., 2018; Zhou et al., 2018b; Xing et al., 2017) similar to copy mechanism (Gu et al., 2016; Merity et al., 2017; He et al., 2017). The controller \( c_t \) is the probability of choosing from knowledge graph entities \( \mathcal{V} \), while \( 1 - c_t \) is the probability of choosing from generic words \( \mathcal{W} \). Note that we take the controller as a special symbol \( KB \) in generic words, so the term \( 1 - c_t \) is already multiplied to \( w_t \). The controller here can be flexibly replaced with any other model.

\[
P\{\{KB, \mathcal{W}\} | y_1 y_2 \ldots y_{t-1}, e(x)\} = \text{softmax}(\phi(d_t)), \quad (3)
\]

\[
w_t = P(\mathcal{V} | y_1 y_2 \ldots y_{t-1}, e(x)), \quad (4)
\]

\[
c_t = P(KB | y_1 y_2 \ldots y_{t-1}, e(x)), \quad (5)
\]

\[
o_t = \{c_t k_t; w_t\}, \quad (6)
\]

where \( \phi \) is the output projection layer, and \( k_t \) is the predicted distribution over knowledge graph entities \( \mathcal{V} \) (detailed in subsection 4.3), and \( o_t \) is the produced distribution over all vocabularies.
4.3 Reasoning Model

To ensure that Qadpt can zero-shot adapt to dynamic knowledge graphs, instead of using attention mechanism on graph embeddings (Ghazvininejad et al., 2018; Zhou et al., 2018b), we leverage the concept of multi-hop reasoning (Lao et al., 2011). The reasoning procedure can be divided into two stages: (1) forming a transition matrix and (2) reasoning multiple hops by a Markov chain.

In the first stage, a transition matrix $T_t$ is viewed as multiplication of a path matrix $R_t$ and the adjacency matrix $A$ of a knowledge graph $\mathcal{K}$. The adjacency matrix is a binary matrix indicating if the relations between two entities exist. The path matrix is a linear transformation of $d_t$, and represents the probability distribution of each head $h \in \mathcal{V}$ choosing each relation type $r \in \mathcal{L}$. Note that a relation type self-loop is added.

$$R_t = \text{softmax}(\theta(d_t)), \quad (7)$$

$$A_{i,j,r} = \begin{cases} 1, & (h_i, r_j, t_\gamma) \in \mathcal{K} \\ 0, & (h_i, r_j, t_\gamma) \notin \mathcal{K} \end{cases}, \quad (8)$$

$$T_t = R_tA, \quad (9)$$

where $R_t \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{L}|}$, $A \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{L}| \times |\mathcal{V}|}$, and $T_t \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$.

In the second stage, a binary vector $s \in \mathbb{R}^{|\mathcal{V}|}$ is computed to indicate whether each knowledge entity exists in the input message $x$. First, the vector $s$ is multiplied by the transition matrix. A new vector $s^t T_t$ is then produced to denote the new probability distribution over knowledge entities after one-hop reasoning. After $N$ times reasoning\(^5\), the final probability distribution over knowledge entities is taken as the generated knowledge entity distribution $k_t$:

$$k_t = s^t (T_t)^N. \quad (10)$$

The loss function is the cross-entropy of the predicted word $o_t$ and the ground-truth distribution:

$$\mathcal{L}(\psi, \phi, \theta) = - \sum_{t=1}^{n} \log o_t(y_t), \quad (11)$$

where $\psi$ is the parameters of GRU layers. Compared to prior work, the proposed reasoning approach explicitly models the knowledge reasoning path, so an updated knowledge graph will definitely change the results without retraining.

\(^5\)We choose $N = 6$ as the maximum length of shortest paths in Figure 3

4.4 Inferring Reasoning Paths

Because this reasoning method is stochastic, we compute the probabilities of the possible reasoning paths by the reasoning model, and infer the one with the largest probability as the retrieved reasoning path.

5 Related Work

The proposed task is motivated by prior knowledge-grounded conversation tasks (Ghazvininejad et al., 2018; Zhou et al., 2018b), but further requires the capability to adapt to dynamic knowledge graphs.

5.1 Knowledge-Grounded Conversations

The recent knowledge-grounded conversation models (Sordoni et al., 2015; Ghazvininejad et al., 2018; Zhu et al., 2017; Zhou et al., 2018b) generated responses conditioned on conversation history and external knowledge. Ghazvininejad et al. (2018) used memory networks (Weston et al., 2015b,a; Sukhbaatar et al., 2015) to attend on external facts, and added the encoded information to the decoding process. Zhu et al. (2017) added a copy mechanism (Gu et al., 2016; Merity et al., 2017; He et al., 2017) for improving its performance. Zhou et al. (2018b) presented two-level graph attention mechanisms (Veličković et al., 2018) to produce more informative responses.

For knowledge from unstructured texts, Ghazvininejad et al. (2018) used bag-of-word representations and Long et al. (2017) applied a convolutional neural network to encode the whole texts. With structured knowledge graphs, Zhu et al. (2017) and Zhou et al. (2018b) utilized graph embedding methods (e.g., TransE (Bordes et al., 2013)) to encode each triplet.

The above methods generated responses without explicit relationship to each external knowledge triplet. Therefore, when a triplet is added or deleted, it is unknown whether their generated responses can change accordingly. Moon et al. (2019) recently presented a similar concept, walking on the knowledge graph, for response generation. Nonetheless, their purpose is to find explainable path on a large-scaled knowledge graph instead of adaptation with the changed knowledge graphs. Hence, the proposed attention-based graph walker may suffer from the same issue as previous embedding-based methods.
Table 4: The results of change rate and accurate change rate.

| Model   | Change Rate | Accurate Change Rate |
|---------|-------------|----------------------|
|         | All Last1   | Last2                | All Last1   | Last2                |
| MemNet  | 92.98       | 31.78                | 37.46       | 62.19                |
| + multi | 98.69       | 77.87                | 81.96       | 62.19                |
| TAware  | 94.38       | 68.33                | 71.86       | 78.23                |
| + multi | 97.74       | 76.68                | 80.50       | 78.23                |
| KAware  | 96.91       | 90.89                | 96.91       | 87.28                |
| + multi | 99.60       | 83.17                | 85.59       | 82.52                |
| Qadpt   | 95.65       | 77.33                | 78.68       | 81.24                |
| + multi | 99.02       | 83.14                | 85.59       | 81.24                |

Table 5: The results of knowledge graph entities prediction.

| Model   | KW Acc | KW/Generic Acc | Generated-KW Acc |
|---------|--------|----------------|-------------------|
| Seq2Seq | 12.10  | 29.08          | 13.30             |
| MemNet  | 22.58  | 39.09          | 39.52             |
| + multi | 35.20  | 54.49          | 60.63             |
| TAware  | 59.71  | 68.61          | 70.18             |
| KAware  | 20.53  | 40.63          | 49.18             |
| Qadpt   | 57.61  | 38.24          | 44.50             |
| + multi | 57.40  | 51.97          | 56.11             |
| + TAware| 56.24  | 53.68          | 58.82             |

5.2 Multi-Hop Reasoning

We leverage multi-hop reasoning (Lao et al., 2011) to allow our model to quickly adapt to dynamic knowledge graphs. Recently, prior work used convolutional neural network (Toutanova et al., 2015), recurrent neural network (Neelakantan et al., 2015; Das et al., 2017), and reinforcement learning (Xiong et al., 2017; Das et al., 2018; Chen et al., 2018; Shen et al., 2018) to model multi-hop reasoning on knowledge graphs, and has proved this concept useful in link prediction. These reasoning models, however, have not yet explored on dialogue generation. The proposed model is the first attempt at adapting conversations via a reasoning procedure.

6 Experiments

For all models, we use gated recurrent unit (GRU) based Seq2Seq models (Cho et al., 2014; Chung et al., 2014; Vinyals and Le, 2015). Both encoder and decoder for HGZHZ are 256 dimension with 1 layer; ones for Friends are 128 dimension with 1 layer.

We benchmark the task, dynamic knowledge-grounded dialogue generation, and corpus DyKgChat by providing a detailed comparison between the prior conversational models and our proposed model as the preliminary experiments. We evaluate their capability of quick adaptation by randomized whole, last 1, last 2 reasoning paths as described in Section 2.1.2. We evaluate the produced responses by sentence-level BLEU-2 (Papineni et al., 2002; Liu et al., 2016), perplexity, distinct-n (Li et al., 2016), and our proposed metrics for predicting knowledge entities descrin section 2.1.1.

Because of the significant data imbalance of Friends, we first train on whole training data, and then fine-tune the models using the subset containing knowledge entities. Early stopping is adopted in all experiments.

6.1 Baselines

We compare our model with prior knowledge-grounded conversation models: the memory network (Ghazvininejad et al., 2018) and knowledge-aware model (KAware) (Zhu et al., 2017; Zhou et al., 2018b). We also leverage the topic-aware model (TAware) (Xing et al., 2017; Wu et al., 2018; Zhou et al., 2018a) by attending on knowledge graphs and using two separate output projection layers (generic words and all knowledge graph entities). In our experiments, MemNet is modified for fair comparison, where the memory pool of MemNet stores TransE embeddings of knowledge triples (Zhou et al., 2018b). The maximum number of the stored triplets are set to the maximum size of all knowledge graphs for
As shown in Table 4, MemNet, TAware and KAware significantly change when the knowledge graphs are largely updated (All) and can also achieve good accurate change rate. For them, the more parts updated (All $>$ $>$ Last2 $>$ Last1), the more changes and accurate changes. However, when the knowledge graphs are slightly updated (Last1 and Last2), the portion of accurate changes over total changes (e.g., the Last1 score 1.17/31.78 for HGZHZ with MemNet model) is significantly low. Among the baselines, KAware has better performance on Last1. On the other hand, Qadpt outperforms all baselines when the knowledge graphs slightly change (Last1 and Last2) in terms of accurate change rate. The proportion of accurate changes over total changes also show significantly better performance than the prior models. Figure 5 shows the distribution of lengths of the inferred relation paths for Qadpt models. After combining TAware or MemNet, the distribution becomes more similar to the test data.

To evaluate the sentence quality, Table 6 presents the BLEU-2 scores (as recommended in the prior work (Liu et al., 2016)), perplexity (PPL), and distinct-n scores. The results show that all models have similar levels of BLEU-2 and PPL, while Qadpt+multi has better distinct-n scores. Even though BLEU-2 is not perfect for measur-

| Model   | HGZHZ          | Friends         |
|---------|----------------|-----------------|
|         | BLEU | PPL | dist-1 | dist-2 | dist-3 | dist-4 | BLEU | PPL | dist-1 | dist-2 | dist-3 | dist-4 |
| Seq2Seq | 14.20 | 94.48 | 0.008  | 0.039  | 0.092  | 0.150  | 15.46 | 73.23 | 0.004  | 0.016  | 0.026  | 0.032  |
| MemNet  | 15.73 | 88.29 | 0.012  | 0.062  | 0.150  | 0.240  | 14.61 | 67.58 | 0.005  | 0.023  | 0.040  | 0.049  |
| + multi | 15.88 | 86.76 | 0.010  | 0.058  | 0.138  | 0.224  | 12.97 | 54.67 | 0.006  | 0.022  | 0.032  | 0.036  |
| TAware  | 15.07 | 81.54 | 0.013  | 0.068  | 0.153  | 0.223  | 14.78 | 60.61 | 0.002  | 0.007  | 0.013  | 0.016  |
| + multi | 13.34 | 80.48 | 0.022  | 0.122  | 0.239  | 0.304  | 15.74 | 56.67 | 0.009  | 0.001  | 0.019  | 0.023  |
| KAware  | 14.14 | 90.11 | 0.011  | 0.061  | 0.135  | 0.198  | 15.70 | 64.70 | 0.002  | 0.009  | 0.017  | 0.021  |
| Qadpt   | 14.52 | 88.24 | 0.013  | 0.081  | 0.169  | 0.242  | 17.01 | 68.27 | 0.002  | 0.008  | 0.013  | 0.016  |
| + multi | 15.47 | 86.65 | 0.021  | 0.129  | 0.259  | 0.342  | 14.79 | 66.70 | 0.005  | 0.023  | 0.041  | 0.051  |
| + TAware| 15.05 | 81.75 | 0.022  | 0.123  | 0.246  | 0.332  | 16.85 | 54.46 | 0.003  | 0.012  | 0.020  | 0.024  |

Table 6: The results of responses generation with BLEU, perplexity (PPL), distinct scores (1-gram to 4-gram).
ing the quality of dialogue generation (Liu et al., 2016), we report it for reference. Overall, the composition of BLEU-2, PPL and distinct-n shows that the quality of the generated responses of these models are not extremely distinguishable.

### 6.3 Human Evaluation

When performing human evaluation, we randomly select examples from the knowledge entities included outputs of all models (therefore this subsection is not used to evaluate generic responses), because it is difficult for human to distinguish which generic response is better (where in most cases, every one is not proper). We recruit fifteen annotators to judge the results. Each annotator was randomly assigned with 20 examples, and was guided to rank the generated responses of the five models: Seq2Seq, MemNet, TAware, KAware, and Qadpt. They were asked to rank the responses according to the following two criteria: (1) fluency and (2) information. Fluency measures which output is more proper as a response to a given input message. Information measures which output contains more correct information (in terms of knowledge words here) according to a given input message and a referred response. The evaluation results are classified into “win”, “tie”, and “lose” for comparison.

The human evaluation results and the annotator agreement in the form of Cohen’s kappa (Cohen, 1960) are reported in Table 7. According to a magnitude guideline (Landis and Koch, 1977), most agreements are substantial (0.6-0.8), while some agreements of Friends are moderate (0.4-0.6). In most cases of Table 7, Qadpt outperforms other four models. However, in Friends, Qadpt, MemNet, and TAware tie closely. The reason might be the lower annotators agreements of Friends, or only the similar trend with automatic evaluation metrics. There are two extra spots. First, Qadpt wins MemNet and TAware less times than winning Seq2Seq and KAware, which aligns with Table 5 and Table 6, where MemNet and TAware show better performance than other baselines. Second, Qadpt wins the baselines more often by fluency than by information, and much more ties happen in the information fields than the fluency fields. This is probably due to the selection of knowledge-contained examples. Hence there is not much difference when seeing the information amount of models. Overall, the human evaluation results can be considered as reference because of the substantial agreement among annotators and the similar trend with automatic evaluation.

### 7 Discussion

The results demonstrate that MemNet, TAware and Qadpt generally perform better than the other two baselines, and they excel at different aspects. While Qadpt generally performs the best in most metrics, TAware shows its power in knowledge graph entities prediction, and MemNet demonstrates its potential in human evaluation.

Specifically for the scenario of zero-shot adaptation, MemNet and TAware show their ability to update responses when the knowledge graphs are largely changed. On the other hand, Qadpt is better to capture minor dynamic changes (Last1 and Last2) and updates the responses according to the new knowledge graphs\(^7\). This demonstrates that Qadpt has the potential to be explore more, while MemNet and TAware attend on the whole graph instead of focusing on the most influential part.

### 8 Conclusion

This paper presents a new task, dynamic knowledge-grounded conversation generation, and a new dataset DyKgChat for evaluation. The dataset is currently provided with a Chinese and an English TV series as well as their correspondent knowledge graphs. This paper also benchmarks the task and dataset by proposing automatic evaluation metrics and baseline models, which can motivate the future research directions.

\(^7\)Examples are given in Appendix.
Acknowledgements

We would like to thank reviewers for their insightful comments. This work was financially supported from the Young Scholar Fellowship Program by Ministry of Science and Technology (MOST) in Taiwan, under Grant 108-2636-E-002-003.

References

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhyenko. 2013. Translating embeddings for modeling multi-relational data. In Advances in neural information processing systems, pages 2787–2795.

Wenhu Chen, Wenhan Xiong, Xifeng Yan, and William Yang Wang. 2018. Variational knowledge graph reasoning. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), volume 1, pages 1823–1832.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 103–111.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. Presented in NIPS 2014 Deep Learning and Representation Learning Workshop.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1):37–46.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zabeer, Luke Vilnis, Ishan Duregukar, Akshay Krishna-murthy, Alex Smola, and Andrew McCallum. 2018. Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning. International Conference on Learning Representations.

Rajarshi Das, Tsendsuren Munkhdalai, Xingdi Yuan, Adam Trischler, and Andrew McCallum. 2019. Building dynamic knowledge graphs from text using machine reading comprehension. In International Conference on Learning Representations.

Rajarshi Das, Arvind Neelakantan, David Belanger, and Andrew McCallum. 2017. Chains of reasoning over entities, relations, and text using recurrent neural networks. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, volume 1, pages 132–141.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, pages 37–49.

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

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 1631–1640.

Yanchao Hao, Yuanzhe Zhang, Kang Liu, Shizhu He, Zhanyi Liu, Hua Wu, and Jun Zhao. 2017. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 221–231.

Shizhu He, Cao Liu, Kang Liu, and Jun Zhao. 2017. Generating natural answers by incorporating copying and retrieving mechanisms in sequence-to-sequence learning. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 199–208.

Pei Ke, Jian Guan, Minlie Huang, and Xiaoyan Zhu. 2018. Generating informative responses with controlled sentence function. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1499–1508.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. Biometrics, pages 159–174.

Ni Lao, Tom Mitchell, and William W Cohen. 2011. Random walk inference and learning in a large scale knowledge base. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 529–539. Association for Computational Linguistics.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. CoNLL 2017, page 333.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In Proceedings of NAACL-HLT, pages 110–119.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016.
How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2122–2132.

Yinong Long, Jianan Wang, Zhen Xu, Zongsheng Wang, Baoxun Wang, and Zhunran Wang. 2017. A knowledge enhanced generative conversational service agent. In Proceedings of the 6th Dialog System Technology Challenges (DSTC6) Workshop.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. Proceedings of the 2017 International Conference on Learning Representations.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subha. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In Proceedings of the 57th Conference of the Association for Computational Linguistics, pages 845–854.

Arvind Neelakantan, Benjamin Roth, and Andrew McCallum. 2015. Compositional vector space models for knowledge base completion. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), volume 1, pages 156–166.

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 on association for computational linguistics, pages 311–318. Association for Computational Linguistics.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Assigning personality/profile to a chatting machine for coherent conversation generation. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, pages 4279–4285. AAAI Press.

Yelong Shen, Jianshu Chen, Po-Sen Huang, Yuqing Guo, and Jianfeng Gao. 2018. Reinforcewalk: Learning to walk in graph with monte carlo tree search. International Conference on Learning Representations.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 196–205.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In Advances in neural information processing systems, pages 2440–2448.

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William Cohen. 2018. Open domain question answering using early fusion of knowledge bases and text. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4231–4242.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Advances in neural information processing systems, pages 3104–3112.

Kristina Toutanova, Danqi Chen, Patrick Pantel, Hofung Poon, Pallavi Choudhury, and Michael Gamon. 2015. Representing text for joint embedding of text and knowledge bases. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1499–1509.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. International Conference on Learning Representation.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. ICML Deep Learning Workshop 2015.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015a. Towards ai-complete question answering: A set of prerequisite toy tasks. arXiv preprint arXiv:1502.05698.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2015b. Memory networks. International Conference on Learning Representations.

Yu Wu, Wei Wu, Dejian Yang, Can Xu, and Zhoujun Li. 2018. Neural response generation with dynamic vocabularies. In Thirty-Second AAAI Conference on Artificial Intelligence.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In Thirty-First AAAI Conference on Artificial Intelligence.

Wenhan Xiong, Thien Hoang, and William Yang Wang. 2017. Deeppath: A reinforcement learning method for knowledge graph reasoning. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 564–573.

Jun Yin, Xin Jiang, Zhengdong Lu, Lifeng Shang, Hang Li, and Xiaoming Li. 2016. Neural generative question answering. In Proceedings of the Workshop on Human-Computer Question Answering, pages 36–42.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018a. Emotional chatting machine: Emotional conversation generation with internal and external memory. In Thirty-Second AAAI Conference on Artificial Intelligence.
Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018b. 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 Qualitative Analysis

Because the models generate different responses with different knowledge given the same input message, it is difficult to design a universal changed knowledge graph for fair comparison of all models. We provide the qualitative analysis for further investigation. Some examples of our proposed Qadpt model with changed knowledge graphs are shown in Table 8. This demonstrates that Qadpt can successfully modify the response with a minor-changed knowledge graph. More examples of response generation are given in Table 9.
甄嬛（Zhen Huan）：怎麼了？——有貓！

浣碧（Wan-Bi）：槿汐，快點去打貓，小姐被貓傷過，見不得貓的！

| Message Input          | Reference Output                                                                 |
|------------------------|----------------------------------------------------------------------------------|
| 雪雁（甘露寺，夜正睡覺，有野貓闖入）：怎麼了？——有貓！ | (Zhen Huan: What’s wrong? – A cat!)                                               |
| 韓碧（一陣慌亂）：槿汐，快點去打貓，小姐被貓傷過，見不得貓的！ | Wan-Bi: Jin-Xi, go and catch the cat. Mistress was hurt by a cat, and became afraid of it. |

| Original KG Path       | Original Output                                                                 |
|------------------------|----------------------------------------------------------------------------------|
|甄嬛(Zhen Huan)        | 主僕(IsServantOrMasterOf) → 槿汐(Jin-Xi)                                       |

| Changed KG Path        | Changed Output                                                                 |
|------------------------|----------------------------------------------------------------------------------|
|甄嬛(Zhen Huan)        | 主僕(IsServantOrMasterOf) → 小允子(Yun)                                         |

| Message Input          | Reference Output                                                                 |
|------------------------|----------------------------------------------------------------------------------|
| [Scene: The restaurant. Chandler and Monica are sitting at a table] | Monica: Oh my God, he's gonna do it now. Please, I cannot watch this, let's go. Chandler: I think we have some time. Have you ever heard him talk? Uh, Phoebe, uh, I would be honoured, uh... Spit it out, David! |
| Original KG Path       | Original Output                                                                 |
|                        | David is, David, is it?                                                          |
| Changed KG Path        | Phoebe → lover → Mike                                                             |
| Changed Output         | Mike?                                                                            |

| Original KG Path       | Original Output                                                                 |
| Phoebe → lover → David | David is, David, is it?                                                          |

| Changed KG Path        | Changed Output                                                                 |
| Phoebe → lover → Mike  | Mike?                                                                            |

| Changed KG Path        | Changed Output                                                                 |
| Phoebe → lover → Mike  | Mike?                                                                            |

Table 8: Some examples of the response generation with a changed knowledge graph.
| Input | Reference |
|-------|-----------|
| 姨娘，您是不是？ (Madame, you are?) | 齊姬，您是喜姬嗎？ (Madame is still the same. Madame is not madame's) |
| 謝韓后（韓后） (Thanks to Consort L) | 謝韓后（韓后） (Thanks to Consort L) |
| 皇上？ (Emperor?) | 謝韓后（韓后） (Thanks to Consort L) |
| 是，韓后說來了。 (Yes, madame said she is coming.) | 是，韓后說來了。 (Yes, madame said she is coming.) |
| 皇上。 (Thanks to the emperor.) | 皇上。 (Thanks to the emperor.) |
| 娘娘給娘娘請安。 (I pay respects to madame.) | 娘娘給娘娘請安。 (I pay respects to madame.) |
| 是，娘娘。您的心思是—— (Sure, madame, you are thinking about —) | 是，娘娘。您的心思是—— (Sure, madame, you are thinking about —) |
| 謝妳。 (Thanks to the emperor.) | 謝妳。 (Thanks to the emperor.) |
| 娘娘，您不習慣嗎？ (Madame, I dare not do it.) | 娘娘，您不習慣嗎？ (Madame, I dare not do it.) |
| 謝皇上。（皇上） (Emperor: It’s okay! They are all imperial doctors.) | 謝皇上。（皇上） (Emperor: It’s okay! They are all imperial doctors.) |
| 皇后（章幹把脈良久）：章太醫，究竟什麼個情形？莫非驚動了胎氣？ | 皇后（章幹把脈良久）：章太醫，究竟什麼個情形？莫非驚動了胎氣？ |
| (Empress: Imperial doctor Zhang, what’s the situation? is it a miscarriage?) | (Empress: Imperial doctor Zhang, what’s the situation? is it a miscarriage?) |
| 江福海：娘娘，奴才怎敢假傳皇上口諭呢？皇上不想見您也不是一日兩日了， | 江福海：娘娘，奴才怎敢假傳皇上口諭呢？皇上不想見您也不是一日兩日了， |
| (Jiang, Fu Hai: Madame, I do not fully you. Dare I say, the emperor do not want to see you.) | (Jiang, Fu Hai: Madame, I do not fully you. Dare I say, the emperor do not want to see you.) |
| 皇帝（眉莊氣息話不聽，甄嬛忙扶住）：無妨！都一樣是太醫。 | 皇帝（眉莊氣息話不聽，甄嬛忙扶住）：無妨！都一樣是太醫。 |
| (Emperor: It’s okay! They are all imperial doctors.) | (Emperor: It’s okay! They are all imperial doctors.) |
| Qadpt Hey, Joey, you’re gonna have a baby. | Qadpt Hey, Joey, you’re gonna have a baby. |
| MemNet Hey, Monica, I’m sorry. | MemNet Hey, Monica, I’m sorry. |
| TAware Hey, Ross. | TAware Hey, Ross. |
| KAware Hey, Ross, you’re a great guy. | KAware Hey, Ross, you’re a great guy. |

Table 9: Some examples of the response generation.