# TopicKA: Generating Commonsense Knowledge-Aware Dialogue Responses Towards the Recommended Topic Fact

**Sixing Wu**[1], **Ying Li**[2*], **Dawei Zhang**[1], **Yang Zhou**[3] and **Zhonghai Wu**[2]
[1]School of Electronics Engineering and Computer Science, Peking University, Beijing, China
[2]National Research Center of Software Engineering, Peking University, Beijing, China
[3]Auburn University, Auburn, Alabama, USA
{wusixing, li.ying, daweizhang}@pku.edu.cn, yangzhou@auburn.edu, zhwu@ss.pku.edu.cn

## Abstract

Insufficient semantic understanding of dialogue always leads to the appearance of generic responses, in generative dialogue systems. Recently, high-quality knowledge bases have been introduced to enhance dialogue understanding, as well as to reduce the prevalence of boring responses. Although such knowledge-aware approaches have shown tremendous potential, they always utilize the knowledge in a black-box fashion. As a result, the generation process is somewhat uncontrollable, and it is also not interpretable. In this paper, we introduce a topic fact-based commonsense knowledge-aware approach, TopicKA. Different from previous works, TopicKA generates responses conditioned not only on the query message but also on a topic fact with an explicit semantic meaning, which also controls the direction of generation. Topic facts are recommended by a recommendation network trained under the Teacher-Student framework. To integrate the recommendation network and the generation network, this paper designs four schemes, which include two non-sampling schemes and two sampling methods. We collected and constructed a large-scale Chinese commonsense knowledge graph. Experimental results on an open Chinese benchmark dataset indicate that our model outperforms baselines in terms of both the objective and the subjective metrics.

## 1 Introduction

Commonsense knowledge plays a crucial role in our daily lives, which consists of a set of widespread and well-known facts: for example, "cats are animals". During conversations, humans can make commonsense inferences that frame their understandings, and help them in making responses [Zhou *et al.*, 2018]. However, unlike human beings, machines can merely utilize the surface knowledge that appears in the given query [Ghazvininejad *et al.*, 2018]. Consequently, traditional models tend to generate generic responses (e.g., "I don't know"), or they generate responses that violate the commonsense (e.g. "The earth is square"). With the maturation of the large-scale commonsense knowledge bases such as ConceptNet [Speer *et al.*, 2017] and ATOMIC [Sap *et al.*, 2019], commonsense knowledge has been successfully applied in the language understanding, generation, and many others.

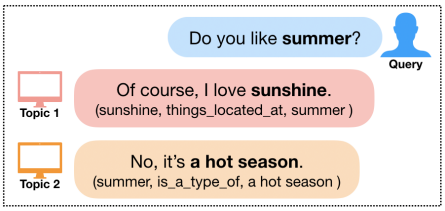


Figure 1: An example to show our TopicKA can generate multiple responses conditioned on different topic facts.

The feasibility of incorporating commonsense knowledge into dialogue systems has already been proved. The first attempt [Young *et al.*, 2018] integrates commonsense knowledge into the retrieval-based dialogue systems. Subsequently, commonsense knowledge has been introduced into the generative models [Zhou *et al.*, 2018]. Moreover, if we broaden our horizons to other types of knowledge-aware dialogue systems [Zhu *et al.*, 2017; Liu *et al.*, 2019], some of them can also be transferred to commonsense knowledge-aware approaches. However, the performance of previous methods is still far from satisfactory. When utilizing the commonsense knowledge, the decision-making with which knowledge facts should be used is usually an uncontrollable black-box process, which always leads to the following issues: (1) Knowledge facts irrelevant to the dialogue context may be used. (2) The generation process is neither controllable nor interpretable. (3) The diversity and the informativeness among the generated responses can not be guaranteed.

This paper proposes a novel **Topic** fact-based commonsense **K**nowledge-**A**ware model, **TopicKA**, for the single-turn open-domain dialogue generation. The most notable difference between our proposal and the previous knowledge-aware approaches is that the dialogue generation framework. As illustrated in Figure 1, instead of generating a response conditioned on the given query message only, as traditional approaches do, the generation in our approach is conditioned on the query message and a controllable and interpretable

*Corresponding author

topic fact simultaneously. Topic facts refer to the commonsense facts that can semantically and logically connect the query message and the target response. Topic facts are recommended by the proposed Topic Fact Recommender, and they are represented as discrete categorical variables, in our approach. Topic facts will be continuously involved in the generation process to guide the direction of the generation. The advantages of using topic facts include the flexibility to control the direction of the generation, higher interpretability to explain the knowledge selection process, and the ability to generate multiple diverse responses.

Three challenges have subsequently emerged. The first is how to select appropriate topic facts. To recommend a topic fact for a query message, we design a recommendation network, Topic Fact Recommender, which recommends topic facts by ranking the preliminary knowledge facts retrieved by entity names. Meanwhile, inspired by the Teacher-Student framework [Hinton *et al.*, 2015], we design a Boost Network that further uses the posterior knowledge as a teacher to improve the ranking performance. The next is how to utilize the recommended topic facts during the response generation. We first propose a diffusion mechanism to diffuse a topic fact, and then the diffused topic fact can guide the generation of the response by using the proposed TFAG (Topic Fact-Aware GRU). The last challenge is how to integrate the recommendation network into the generation network. The generation network accepts a discrete categorical topic fact as an input, which requires sampling a topic fact from the output of the recommendation network. The sampling process would make the integration non-differentiable. To overcome this, we propose two non-sampling integration schemes: Two-Stage Learning and Multi-Task Learning; and we propose two differentiable sampling approaches: one is inspired by [Zhao *et al.*, 2017], and another one approximates such a sampling process by using Gumbel-Softmax [Jang *et al.*, 2017]. Both the automatic evaluation and the human annotation on an open Chinese benchmark dataset indicate that the proposed TopicKA can outperform the SOTA approach CCM [Zhou *et al.*, 2018] in terms of most metrics. In the case study, we verify the controllability, interpretability, and diversity.

Our contributions are: (1) We proposed a novel topic fact-based commonsense knowledge-aware approach, TopicKA. (2) We proposed four schemes to integrate the generator and the recommender. (3) We collected and constructed a large-scale Chinese commonsense knowledge graph. Experimental results indicate our TopicKA outperforms various kinds of baselines.

## 2 Approach

### 2.1 Problem Formulation

Let $X$ denote the query message, $Y$ denote the target response, $F = \{f\}$ denote a set of facts retrieved from the commonsense knowledge graph $\mathcal{G}$, and $f_t$ denote the topic fact. Each fact is organized as a SVO (Subject, Verb, Object) triplet, namely, $f = (e_1, r, e_2)$. Instead of modeling dialogue generation as $P(Y|X)$, TopicKA generates a response conditioned on the query message and a recommended topic fact simultaneously; that is, $P(Y|X, f_t)$.

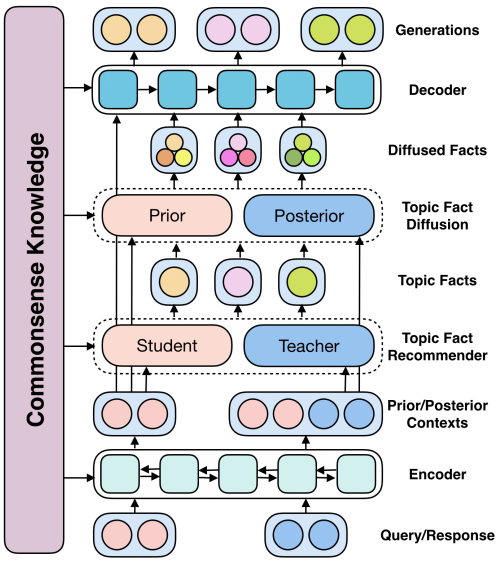


Figure 2: An overview of TopicKA. Shapes in pink are prior knowledge or modules, shapes in blue are posterior knowledge or modules.

### 2.2 Context Encoder

Context Encoder $C_\theta(\cdot)$ is a bi-directional GRU network [Cho *et al.*, 2014], which encodes an utterance $U$ into contextual intermediate representations $\mathbf{H^u} = (\mathbf{h_1^u}, \ldots, \mathbf{h_T^u})$. The forward GRU reads the $U$ in the normal order; the backward GRU reads the $U$ in the inversed order. At the time step $t$, the output $\mathbf{h_t^u}$ is the concatenation of two directions' outputs.

### 2.3 Topic Fact Recommender

The topic fact $f_t$ is selected from the retrieved knowledge facts $F = \{f\}$ given the prior context $\mathbf{H^x}$. The probability that $f_i$ can be recommended as a topic fact is given by a probability distribution over the $F$:

$$p_t(f_i) = \frac{\exp(\mathbf{q_i^x} \cdot \mathbf{f_i})}{\sum_{j=1:|F|} \exp(\mathbf{q_i^x} \cdot \mathbf{f_j})}$$

$$\mathbf{q_i^x} = \sum_{t=1:|X|} a_t \mathbf{h_t^x} \mathbf{W_v^a}, a_t = \frac{\exp\left(\alpha(\mathbf{f_i}, \mathbf{h_t^x} \mathbf{W_k^a})\right)}{\sum_{j=1:|X|} \exp\left(\alpha(\mathbf{f_j}, \mathbf{h_t^x} \mathbf{W_k^a})\right)} \quad (1)$$

where $\mathbf{f}$ is the embedding of $f$, $\mathbf{q_i^x}$ is an attentive context, and $\alpha(\cdot)$ is an alignment function [Luong *et al.*, 2015], which measures the relevance between the two inputs.

Inspired by the success of distillation learning (Teacher-Student framework) [Hinton *et al.*, 2015], we design a **Boost Network** (teacher) to improve the ranking performance of the recommendation network (student), by applying posterior soft labels. The Boost Network has a network structure similar to that of the student network, but it additionally employs the posterior context $\mathbf{H^y}$:

$$p_t^t(f_i) = \frac{\exp(\mathbf{q^y} \cdot \mathbf{f_i})}{\sum_{j=1:|F|} \exp(\mathbf{q^y} \cdot \mathbf{f_j})}, \mathbf{q^y} = [\mathbf{h_{|X|}^x}; \mathbf{h_{|Y|}^y}] \mathbf{W_k^t} \quad (2)$$

Two networks share the same Context Encoder and the same embedding. Importantly, two networks are jointly trained; therefore, the objective is to minimize:

$$\mathcal{L}_{\phi+\theta} = CE(\hat{f}_t, p_t) + CE(\hat{f}_t, p_t^t) + KL(p_t^t || p_t) \quad (3)$$

where $\phi$ denotes the parameters of two networks, $\hat{f}_t$ is the labeled topic fact, $CE(\cdot,\cdot)$ is the Cross-Entropy that is used to optimize two networks, and $KL(\cdot||\cdot)$ is the Kullback–Leibler Divergence [Kullback and Leibler, 1951] to force the student to generate a distribution similar to the teacher's distribution.

## 2.4 Knowledge-Aware Dialogue Generation

**Topic Fact Diffusion Network**

A single fact is usually not enough to cover the knowledge of a topic. Therefore, we design a Topic Fact Diffusion Network to retrieve more relevant facts based on the $f_t$. Given the context and a topic fact $f_t$, the diffusion network outputs a probability distribution $p_z$ over the $F$, by using an attention function $D$ with softmax:

$$
\begin{aligned}
D(\mathbf{Query}) &= softmax(elu(\mathbf{F_{sv}W_{ft}})elu(\mathbf{Query}^\top)) \\
p_z = D([\mathbf{f_{t,sv}};\mathbf{h^x_n}]\mathbf{W^z_x}), p_z^{po} &= D([\mathbf{f_{t,sv}};\mathbf{h^x_n};\mathbf{h^y_m}]\mathbf{W^z_{xy}}) \\
\mathbf{f_z} &= \textstyle\sum p_z(f_i)\mathbf{f_i} \text{ if test else } \sum \mathrm{p_z^{po}(f_i)}\mathbf{f_i}
\end{aligned} \tag{4}
$$

where $elu$ is an activation function [Clevert *et al.*, 2016], $p_z[i]$ is regarded as the relevance of the $i$-th fact to the central topic fact $f_t$, and the fused $\mathbf{f_z}$ is the diffused topic fact embedding. To improve the diversity, the objective entity will not be considered during generating $p_z$, $\mathbf{F_{sv}}/\mathbf{f_{c,sv}}$ is the corresponding ablation of $\mathbf{F}/\mathbf{f_t}$. As the recommendation network, the diffusion network utilizes the posterior context during the traning , where $p_z$ is the prior one, $p_z^{po}$ is the posterior one. $KL(p_z^{po}||p_z)$ is subsequently adopted.

**Topic Fact-Aware GRU**

Intuitively, a straightforward approach, by which to use the introduced topic facts to control the generation, is directly attaching such contents to the input of GRU. However, this approach usually fails to generate a fluent and diverse response, because it lacks enough flexibility to control the degree of participation of the introduced topic facts. Inspired by [Arevalo *et al.*, 2017; Yao *et al.*, 2017], we propose a **TFAG** (Topic Fact-Aware GRU) to integrate the diffused topic fact flexibly. A TFAG unit consists of two different GRUs that accept different inputs, and a $FusionUnit$ ( see [Arevalo *et al.*, 2017] for the detail) to fuse two GRUs' internal states. At the time step $t$, the internal state $\mathbf{s_t}$ of TFAG can be updated as :

$$
\begin{aligned}
\mathbf{s^c_t} &= GRU^s(\mathbf{s_{t-1}},\mathbf{c_t},\mathbf{y_{t-1}}) \\
\mathbf{s^k_t} &= GRU^k(\mathbf{s_{t-1}},\mathbf{c_t},\mathbf{f_z},\mathbf{c^f_t},\mathbf{y_{t-1}}) \\
\mathbf{s_t} &= o_t \times \mathbf{s^c_t} + (1-o_t)\mathbf{s^k_t}\ ,\ o_t = FusionGate(\mathbf{s^c_t},\mathbf{s^k_t})
\end{aligned} \tag{5}
$$

where the first standard $GRU^s$ works as most conventional GRUs do, which outputs its own state $\mathbf{s^c_t}$ given the attention $\mathbf{c_{t-1}}$ over the query message $X$ and the embedding of the last generated word $\mathbf{y_{t-1}}$. The next GRU is the knowledge-aware $GRU^k$, which additionally accepts the diffused topic fact embedding $\mathbf{f_z}$, and the attention $\mathbf{c^f_t}$ over the retrieved knowledge facts $F$ as inputs. Subsequently, two states $\mathbf{s^c_t}$ and $\mathbf{s^k_t}$ can be fused by adopting the fusion gate $o_t$.

Both the attention $\mathbf{c_t}$ over the query message $X$, and the attention $\mathbf{c^f_t}$ over the retrieved knowledge facts, can be computed by using the approach proposed by [Luong *et al.*, 2015].

**Response Generation**

Following [Wu *et al.*, 2020], the next target word $y_t$ can be generated by choosing a word from the fixed vocabulary $V$, adopting an entity $\in F$, or copying a word from the query message $X$; therefore, the corresponding probability distribution $p(y_t)$ is calculated as:

$$
\begin{aligned}
p(y_t) &= z_t^V \cdot p_V(y_t) + z_t^K \cdot p_K(y_t) + z_t^C \cdot p_C(y_t) \\
[z_t^V, z_t^K, z_t^C] &= \mathrm{softmax}([\mathbf{s_t};\mathbf{c_t};\mathbf{y_{t-1}}]\mathbf{W_o})
\end{aligned} \tag{6}
$$

where $p_V(y_t), p_K(y_t), p_C(y_t)$ are the distributions over the fixed vocabulary $V$, the retrieved knowledge facts $F$, and the query message $X$, respectively; $z_t^V, z_t^K, z_t^C$ are three gates to control the contribution of three types of words. Such three distributions can be calculated by:

$$
\begin{aligned}
p_V &= softmax(MLP_2(\mathbf{s_t};\mathbf{c_t};\mathbf{y_{t-1}})) \\
p_K &= softmax(elu(\mathbf{FW_f})\, elu([\mathbf{s_t};\mathbf{y_{t-1}};\mathbf{f_z}]\mathbf{W_d})^\top) \\
p_C &= softmax(elu(\mathbf{H^xW_{cs}})\, elu([\mathbf{s_t};\mathbf{c_t};\mathbf{y_{t-1}}]\mathbf{W_{ct}})^\top)
\end{aligned} \tag{7}
$$

where $MLP_2$ refers to a 2-layer MLP network. Subsequently, the objective of the generation network is to minimize:

$$
\begin{aligned}
\mathcal{L}_{\pi+\theta} \quad = \quad & -\textstyle\sum_i^{|Y|} log(p(y_i|y_{<i},X,f_t,F)) \\
& + KL(p_z^{po}||p_z) + \mathcal{L}_{bow} + \mathcal{L}_{mode}
\end{aligned} \tag{8}
$$

where $\pi$ denotes the parameters of the generation network, $\mathcal{L}_{bow}$ is the bag-of-words loss to improve the fluency and avoid KL annealing (see [Zhao *et al.*, 2017]), $\mathcal{L}_{mode}$ (see [Zhou *et al.*, 2018]) is the teach-force loss to help the generation network to become more accurate when selecting a target word from three types of words.

## 2.5 Integration Schemes

The generation network requires a discrete categorical topic fact as an input, which means we need to sample a topic fact from the distribution outputted by the recommendation network, i.e., $f_t^* = \arg\max_{f_t} p_c(f_t)$. Considering such sampling is a nondifferentiable process, we here propose following schemes to overcome this issue:

**Two-Stage (TopicKA$_{ts}$)**

We separately train the recommendation network and the generation network in two different stages. During the inference, the recommendation network is used firstly to predict topic facts, and then the generation network can generate responses with the previously predicted topic facts. The topic fact label for the training is constructed under the distant supervision assumption; for a fact $f_i = (e_1, r, e_2) \in F$, if its subjective entity $e_1$ can be matched in the query $X$, and its objective entity $e_2$ can be matched in the response $Y$, then $f_i$ is distantly labeled as a topic fact.

**Multi-Task (TopicKA$_{mt}$)**
Considering that two networks use the same Context Encoder, they can be jointly trained by sharing the parameters of Context Encoder $\theta$. Except for the training, the distantly labeled topic facts and the two-stage inference are kept. The objective is $\mathcal{L}_{MT} = \alpha * \mathcal{L}_{\phi+\theta} + \mathcal{L}_{\pi+\theta}$, where $\alpha$ it s a hyper parameter to coordinate the training speed of two networks, in our experiment, $\alpha = 0.25$.

**Latent Variable (TopicKA$_{lv}$)**
Inspired by the CVAE-based approaches [Zhao *et al.*, 2017], topic facts can be regarded as latent variables. Consequently, we can also optimize the corresponding lower bound:

$$\mathcal{L}_{Latent} = \mathbb{E}_{f_t \sim p_t^t}[\mathcal{L}_{\pi+\theta,f_t}] + KL(p_t^t||p_t) \tag{9}$$

where the exception $\mathbb{E}_{f_t \sim p_t^t}[\mathcal{L}_{\pi+\theta,f_t}]$ can be approximated by sampling $N$ times $f_t$ and then calculating the average: $\frac{1}{N}\sum \mathcal{L}_{\pi+\theta}$.

**Gumbel-Softmax (TopicKA$_{gs}$)**
Topic facts are categorical variables. Thus, we adopt the Gumbel-Softmax [Jang *et al.*, 2017] to approximate the categorical sampling in a differentiable way. The Gumbel-Softmax distribution is given by:

$$p_g[i] = \frac{\exp\left((\log(p_t^t[i]) + g_i)/\omega\right)}{\sum_j \exp\left((\log(p_t^t[j]) + g_j)/\omega\right)} \tag{10}$$

where $\omega = 0.1$ is a temperature value; $g = -\log(-\log(u))$; and $u$ is sampled from the distribution $Uniform(0, 1)$.

# 3 Experiments

## 3.1 Settings

**Dataset**
Our approach is evaluated on an open Chinese benchmark dataset [Li and Yan, 2018], which is collected from the largest Chinese SNS (weibo.com). We collect the commonsense knowledge from the ConceptNet (conceptnet.io). We align the dialogues and the knowledge facts by names. For example, if there is a word 'cat' in the dialogue, then facts whose subject/object entity is 'cat' are aligned. After the alignment, the remaining aligned data are randomly divided into three sets: training set, validation set, and test set, which have 847K, 30K, and 30K pairs of dialogue, respectively. Besides, the crawled commonsense knowledge graph includes 27K entities, 26 relations, and 661K facts, respectively . The entity/relation embedding of the knowledge graph is learned by using TransE [Bordes *et al.*, 2013]. The vocabulary size is set to 50,000. Other OOV words are replaced by $unk$. Our experimental resources are open released[1].

**Comparison Models**
We select five approaches as our baselines. **ATS2S** : The Attentive Seq2Seq [Luong *et al.*, 2015]. **MMI**: The bidi-MMI [Li *et al.*, 2016a], which first uses $p(y|x)$ to generate 10 responses for each query using beam-search, then applies the inversed model $p_{inv}(x|y)$ to re-rank. **HGFU**: HGFU uses cue words predicted by PMI to control the generation [Yao *et al.*, 2017]. **GenDS**: GenDS is able to copy entities during the generation [Zhu *et al.*, 2017]. **CCM**: The SOTA model in the commonsense knowledge-aware dialogue generation [Zhou *et al.*, 2018]. In the experiments, CCM is evaluated with the official code, while other models are re-implemented by Tensorflow. Most hyper parameters are selected from the CCM: batch size is 50, word embedding dimension is 300, GRU dimension is 512, Adam with the initial learning rate 0.0001 is used to optimize, and the maximum epoch is limited to 20. For the TopicKA$_{lv/gs}$, their parameters are initialized from the TopicKA$_{mt}$ for the faster convergence. Meanwhile, considering the uncertainty from random sampling, we have two sampling strategies for TopicKA$_{lv/gs}$ in the inference stage. The first strategy (TopicKA$^{max}_{lv/gs}$) selects the fact with the maximum probability. The second strategy employs the MMI (TopicKA$^{mmi}_{lv/gs}$). We first generate 10 responses $y_{1:10}$ with the sampled 10 facts, and then we select the best candidate via $y^* = \arg\max_{y_i} p_{inv}(x|y_i)$.

**Metric**
The first type focuses on knowledge utilization. **EntN** is the number of generated entities per generation [Zhou *et al.*, 2018], **EntR** is the ratio of the recalled entities. The second type reveals the relevance between the generation and the ground-truth. Following [Liu *et al.*, 2016], we adopt two embedding based metrics, **EmbedA** and **EmbedX** , which consider the averaged embedding and the extreme value of each dimension, respectively. And word overlap-based **ROUGE** and **BLEU1/2**. To measure the diversity, we report the ratio of distinct uni/bi-grams (**DIST1/2**) in all generated words [Li *et al.*, 2016a]. Lastly, we use the **Entropy** to measure the informativeness [Mou *et al.*, 2016].

## 3.2 Experimental Results

The left block of Table 1 shows the automatic evaluation results. In short, TopicKAs are better than baselines in terms of most metrics. Although our TopicKAs (except TopicKA$_{mt}$) outperform baselines in terms of BLEU2, the non-zero cases (only 16%) cannot support the significance. Our approach has outstanding advantages in generating knowledgeable (EntN/R), diverse (DIST1/2) and informative (Entropy) responses. It can be attributed to (1) the effective utilization of the knowledge; (2) the introduction of the topic fact; (3) the design of the network architecture. In the relevance part, our approach's advantage is also notable except on the BLEU2. The mediocre performance on the BLEU2 is that BLEU-N itself only simply measures the n-gram overlap without considering the semantic meaning. Meanwhile, researchers have shown the low relevance between the BLEU and the human annotation. [Liu *et al.*, 2016].

**The Difference among Schemes**
In the first two non-sampling schemes, the multitask scheme TopicKA$_{mt}$ slightly outperforms TopicKA$_{ts}$, which shows that the recommendation network and the generation network can benefit from each other by sharing the Context Encoder. In the comparison of the next two sampling schemes, Gumbel-Softmax brings more diversity and informativeness,

[1] https://github.com/pku-orangecat/IJCAI2020-TopicKA

| Metric | EntN | EntR | EmbedA | EmbedX | ROUGE | BLEU1 | BLEU2 | DIST1 | DIST2 | Entropy | $Ap_{win}$ | $Ap_{tie}$ | $Ap_{lose}$ | $In_{win}$ | $In_{tie}$ | $In_{lose}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ATS2S | 0.59 | 0.15 | 0.777 | 0.530 | 9.95 | 9.55 | 2.92 | 0.79 | 3.12 | 6.47 | **56.5%** | 3.5% | 40.0% | **72.8%** | 3.4% | 23.8% |
| MMI | 0.74 | 0.18 | 0.791 | 0.562 | 10.57 | 11.87 | 3.98 | 1.59 | 7.42 | 7.64 | **60.7%** | 1.2% | 38.1% | **69.5%** | 0.7% | 29.8% |
| HGFU | 0.57 | 0.13 | 0.787 | 0.521 | 7.57 | 6.57 | 2.01 | 2.32 | 7.47 | 7.68 | **64.7%** | 2.7% | 33.6% | **62.0%** | 2.2% | 35.8% |
| GenDS | 0.97 | 0.37 | 0.792 | 0.542 | 12.82 | 11.03 | 3.42 | 0.92 | 4.27 | 6.31 | **74.7%** | 2.8% | 22.5% | **79.3%** | 2.7% | 18.0% |
| CCM | 1.09 | 0.37 | 0.798 | 0.544 | 13.46 | 14.38 | 4.75 | 1.15 | 4.87 | 6.49 | **60.7%** | 1.2% | 38.1% | **82%** | 0.8% | 17.2% |
| $\text{Topic}_{ts}$ | **1.69** | **0.44** | **0.807** | **0.565** | **13.36*** | **15.67** | 4.86 | **4.27** | **22.83** | **8.75** | - | - | - | - | - | - |
| $\text{Topic}_{mt}$ | **1.74** | **0.44** | **0.805** | **0.581** | **13.60** | **15.09** | 4.63 | **4.21** | **24.93** | **8.85** | - | - | - | - | - | - |
| $\text{Topic}_{lv}^{max}$ | **1.74** | **0.46** | **0.816** | **0.570** | **14.12** | **16.09** | 4.99 | **3.95** | **22.47** | **8.68** | - | - | - | - | - | - |
| $\text{Topic}_{gs}^{max}$ | **1.68** | **0.45** | **0.815** | **0.595** | **13.78** | **15.73** | 4.90 | **4.51** | **24.87** | **8.82** | - | - | - | - | - | - |
| $\text{Topic}_{lv}^{mmi}$ | **1.81** | **0.47** | **0.819** | **0.582** | **14.26** | **16.31** | 5.07 | **4.05** | **23.56** | **8.73** | - | - | - | - | - | - |
| $\text{Topic}_{gs}^{max}$ | **1.75** | **0.46** | **0.812** | **0.565** | **13.90** | **15.94** | 4.99 | **4.59** | **26.09** | **8.88** | - | - | - | - | - | - |

Table 1: The left block shows the automatic evaluation results. The right block shows human annotation results. **Ap** is Appropriateness, **In** is Informativeness. **Scores** in bold means the corresponding TopicKA is significantly better than all baselines (sign test, p-value $< 0.005$, ties are removed). *: Although CCM outperforms $\text{Topic}_{ts}$ in terms of the averaged ROUGE, the case number that $\text{Topic}_{ts}$ wins CCM is more than the number that CCM wins $\text{Topic}_{ts}$, therefore, $\text{Topic}_{ts}$ is better than CCM in the sign-test.

while Latent-Variable brings higher knowledge utilization rate and relevance. After applying the re-ranking, the overall performance of two schemes increases, indicating the effectiveness of re-ranking. Although the performance gap between the non-sampling schemes and the sampling schemes is not quite notable in the automatic evaluation, they are very different in actual usage. The advantage of sampling schemes is generating multiple different responses given the same query message. We will exhibit this in the case study.

### Human Annotation

We invited three volunteers to annotate the generated responses. We randomly sampled 200 queries from the test set, and then conducted the pair-wise comparison ($\text{TopicKA}_{lv}^{mmi}$ vs. baselines, 1000 comparisons in total). Volunteers were required to judge which response is better (ties should be avoided as possible), in terms of two metrics: (1) Appropriateness (the fluency and the relevance to the query); (2) Informativeness (how much knowledge can be provided). We counted the agreement among the volunteers. For the appropriateness, 2/3 agreement (i.e., the percentage of cases that at least 2 volunteers gave the same label) is 97%, and the 3/3 agreement is 57%. For the informativeness, 2/3 agreement is 98%, and the 3/3 agreement is 63%. As shown in the right block of Table 1, our approach is better than the baselines in terms of both metrics, which indicates the generated responses of TopicKA are more acceptable by humans. We noticed that the performance of GenDS and CCM on the human annotation is even worse than the results on the automatic evaluation. This is because the generated entity words are a little unnatural (i.e., the generation is not fluent).

### Ablation Study

The ablation results have been reported in Table 2. **Firstly**, we focus on the ranking performance of the recommendation network using Hit@1(the ratio that the labeled topic fact is the first rank fact, higher is better) and MeanRank (the averaged rank of the labeled topic fact, lower is better), $\text{TopicKA}_{mt}$ is the baseline (w/o None). Comparing the joint training (w/o None) and the two-stage (w/o Multitask) training, it can be verified that two networks can promote each other. Meanwhile, the results also demonstrate that

| Rec. w/o | None | Multitask | Boost+Multitask | - |
|---|---|---|---|---|
| MeanRank | 4.99 | 5.12 | 5.20 | - |
| Hit@1 | 32.0% | 31.0% | 31.0% | - |
| Gen. w/o | None | Diff | TFAG+Diff | Topic Fact |
| EntN | 1.69 | 1.17 | 1.18 | 1.23 |
| Dist2 | 22.83 | 14.40 | 13.90 | 11.98 |
| Entropy | 8.75 | 7.96 | 7.92 | 7.52 |

Table 2: The upper block shows ablation results for the recommendation network. The bottom block shows the ablation results for the generation network.

the teacher Boost network can improve the ranking performance (MeanRank, w/o Multitask vs. w/o Boost+Multitask). **Secondly**, we also ablate the techniques related to the topic fact, $\text{TopicKA}_{ts}$ is the baseline (w/o None). After the ablation of the topic fact diffusion (w/o Diff), the performance sharply decreases, which indicates the diffusion mechanism can effectively supplement background knowledge. Next, we further remove the TFAG (w/o TFAG+Diff), the overall performance decreases too. It shows that TFAG can bring a more flexible way to unitize the topic facts. Finally, we completely remove the participation of topic facts (w/o Topic Fact, TFAG, and Diff are also removed). Without the condition of the topic fact, the model can be more freedom in utilizing the knowledge (EntN), but the generated response will be more general and boring (Dist2 and Entropy).

## 3.3 Case Study

Table 3 reports four cases. In case 1, we compare our approach ($\text{TopicKA}_{lv}^{mmi}$) with baselines. It can be seen that all baseline models have generated responses that are irrelevant to the query, the response generated by our approach is not only relevant but also correctly shows the commonsense knowledge that the cloudy day is not hot. In case 2/3, we show that our approach can generate multiple diverse responses conditioned on the different recommended topic facts. Such generated responses are relevant to both query message and the corresponding recommended topic facts, which indicates two things: 1) TopicKA can control the direction by using different topic facts. 2) The motivation why TopicKA generates a specific response is some-

| |
|---|
| #1: 天气预报说明天阴天.<br>The forecast says tomorrow is cloudy. |
| ATS2S: 我也是，今天出太阳了。 Me too. It's sunny today.<br>MMI: 我也是，天天下雨。Me too, it rains every day.<br>HGFU: 我也是， 今天下雨。Me too. It's raining today<br>GenDS: 我也是 ，今天大太阳。 Me too. It's sunny today.<br>CCM: 我明天也要去。Tomorrow, I will go too.<br>TopicKA: 明天还不热的呢？ Is it not hot tomorrow? |
| #2: 是呀,这边的天空真的**好看**！<br>Yeah, the sky here is **beautiful**. |
| TopicKA-Topic1: 好**喜欢**这种类型的天空. (好看,Causes,喜欢)<br>I **love** this kind of sky. (beautiful,Causes,love)<br>TopicKA-Topic2: 嗯,空气好**美**的地方. (好看,Synonym, 美)<br>Uh-huh, a place with **pretty** air. (beautiful,Synonym, pretty) |
| #3: **冰淇淋**好吃！<br>**Ice cream** is delicious! |
| TopicKA-Topic1: 超爱吃！(冰淇淋,IsA,食物)<br>I like it very much! (ice cream, IsA, food)<br>TopicKA-Topic2: 冰淇淋也是绝了. (冰淇淋,HasProperty, 好吃)<br>Ice cream is excellent. (ice cream,HasProperty, delicous) |
| #4(Bad Case): 你是**学**的什么专业?<br>What major do you **learn**? |
| TopicKA: 我学**比赛**，在培训。（比赛,Causes,学）<br>I learn the **match**, I'm training now. (match, Causes, learn) |

Table 3: Case Study, TopicKA refers to TopicKA$^{lv}_{mmi}$.

what interpretable, because the used topic fact has the human-readable semantic meaning. Meanwhile, topic facts can be either explicitly used in the generated response (Case 2), or implicitly used (Case 3). The last case is a bad case, which indicates what we need to improve in the future. In our approach, if the recommended topic fact is distorted, then the generated response may go in a strange direction.

## 4 Related Work

**Open-Domain Dialogue Generation**

The success of the Seq2Seq learning [Sutskever *et al.*, 2014] has motivated the investigation of dialogue systems. In Seq2Seq models, the given query message is first encoded into contextual vectors by an Encoder, and then a Decoder uses such vectors to generate a response. However, Seq2Seq models tend to generate meaningless responses. Namely, some high-frequency general patterns are easy to be generated, in spite of different queries. Researchers have tried to overcome this issue from multiple aspects; for example, replacing the conventional training objective[Li *et al.*, 2016a]; improving the decoding process [Li *et al.*, 2016b].

**Knowledge-Aware Models**

Unlike human beings, who can associate the background knowledge from the mind, traditional models can merely access very limited information from the plain text of a given query [Ghazvininejad *et al.*, 2018]. To tackle this issue, researchers begin to introduce knowledge bases into dialogue systems. [Zhu *et al.*, 2017] allows a model to deal with OOV knowledge entities. [Young *et al.*, 2018] enhances dialogue retrieval with commonsense knowledge; [Liu *et al.*, 2019] augments the structured knowledge base with unstructured knowledge bases; In our task, dialogue generation with commonsense knowledge; the SOTA approach [Zhou *et al.*, 2018] designs graph attention mechanisms to utilize commonsense knowledge. Compared with such approaches, TopicKA can exhibit the central topic fact of a generated response, and TopicKA is controllable such that TopicKA can generate multiple diverse responses based on different topic facts.

**Content-Introducing Generation**

Cue words selected by the external Pointwise Mutual Information (PMI) are first used to control the generation [Mou *et al.*, 2016] explicitly. Then the approach has been improved to a more flexible implicit way [Yao *et al.*, 2017]. However, PMI-based approaches must inefficiently search the whole vocabulary space to find a cue word during the inference. Meanwhile, the searched cue words are usually irrelevant to the dialogue context. Unlike the previous two works, contents can also be sampled from a latent space. The latent space can be either a fixed dialogue action space [Zhao *et al.*, 2018] or a fixed vocabulary space [Gao *et al.*, 2019b; Gao *et al.*, 2019a]. TopicKA is also different from such approaches. First, the content that is introduced in the TopicKA is the high-quality commonsense knowledge facts, which are more reasonable/reliable than actions or words. Next, unlike previous approaches, which sample contents from a limited content space, our content space is quite larger than theirs. The largest content space of previously mentioned approaches is $\mathbb{R}^{|V|}$ (vocabulary space), but ours is up to $\mathbb{R}^{|E|\times|E|\times|R|}$, where $|E|/|R|$ denotes the number of the entities/relations.

## 5 Conclusion and Future Work

To bridge the gap of the knowledge between the machine and the human in the context of dialogue response generation, this paper proposes a novel flexible model TopicKA, which can utilize the commonsense knowledge. TopicKA can generate diverse and informative responses conditioned on an interpretable and a controllable topic fact. To recommend topic facts, we propose a teacher-student recommendation network. We propose four schemes to integrate the generation network and the recommendation. Meanwhile, we collect and construct a Chinese commonsense knowledge graph. Experimental results on both objective evaluation and subjective evaluation, indicating our TopicKA can significantly outperform the state-of-the-art CCM in terms of most metrics.

However, there is still much room for improvement in the future. First, the Topic Fact Recommender can be further improved. Second, although the commonsense knowledge can significantly enrich the background knowledge understanding/generating, it is still far away from the satisfactory. In the future, we will continue to introduce more types of knowledge. Lastly, we will continue to improve the diversity, interpretability, and controllability of knowledge-aware dialogue generation models.

## Acknowledgments

This work is supported by the National Key R&D Program of China (Grant No. 2017YFB1002000), and PKU-Tencent Joint Innovation Research Program.

## References


[Arevalo *et al.*, 2017] John Arevalo, Thamar Solorio, Manuel Montes y Gómez, and Fabio A. González. Gated multimodal units for information fusion, 2017.

[Bordes *et al.*, 2013] Antoine Bordes, Nicolas Usunier, Jason Weston, and Oksana Yakhnenko. Translating Embeddings for Modeling Multi-Relational Data. In *NIPS*, pages 2787–2795, 2013.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, pages 1724–1734, 2014.

[Clevert *et al.*, 2016] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). In *ICLR*, 2016.

[Gao *et al.*, 2019a] Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, and Shuming Shi. Generating multiple diverse responses for short-text conversation. In *AAAI*, pages 6383–6390, 2019.

[Gao *et al.*, 2019b] Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, Guodong Zhou, and Shuming Shi. A discrete CVAE for response generation on short-text conversation. In *EMNLP*, pages 1898–1908, 2019.

[Ghazvininejad *et al.*, 2018] Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. A knowledge-grounded neural conversation model. In *AAAI*, pages 5110–5117, 2018.

[Hinton *et al.*, 2015] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

[Jang *et al.*, 2017] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2017.

[Kullback and Leibler, 1951] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[Li and Yan, 2018] Juntao Li and Rui Yan. Overview of the NLPCC 2018 shared task: Multi-turn human-computer conversations. In *NLPCC*, pages 446–451, 2018.

[Li *et al.*, 2016a] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *NAACL*, pages 110–119, 2016.

[Li *et al.*, 2016b] Jiwei Li, Will Monroe, and Dan Jurafsky. A simple, fast diverse decoding algorithm for neural generation. *CoRR*, abs/1611.08562, 2016.

[Liu *et al.*, 2016] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*, pages 2122–2132, 2016.

[Liu *et al.*, 2019] Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. Knowledge Aware Conversation Generation with Explainable Reasoning over Augmented Graphs. In *EMNLP*, pages 1782–1792, 2019.

[Luong *et al.*, 2015] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, pages 1412–1421, 2015.

[Mou *et al.*, 2016] Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *COLING*, pages 3349–3358, 2016.

[Sap *et al.*, 2019] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *AAAI*, pages 3027–3035, 2019.

[Speer *et al.*, 2017] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, pages 4444–4451, 2017.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112, 2014.

[Wu *et al.*, 2020] Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In *ACL*, 2020.

[Yao *et al.*, 2017] Lili Yao, Yaoyuan Zhang, Yansong Feng, Dongyan Zhao, and Rui Yan. Towards implicit content-introducing for generative short-text conversation systems. In *EMNLP*, pages 2190–2199, 2017.

[Young *et al.*, 2018] Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. Augmenting end-to-end dialogue systems with commonsense knowledge. In *AAAI*, pages 4970–4977, 2018.

[Zhao *et al.*, 2017] Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL*, pages 654–664, 2017.

[Zhao *et al.*, 2018] Tiancheng Zhao, Kyusong Lee, and Maxine Eskénazi. Unsupervised discrete sentence representation learning for interpretable neural dialog generation. In *ACL*, pages 1098–1107, 2018.

[Zhou *et al.*, 2018] Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. Commonsense Knowledge Aware Conversation Generation with Graph Attention. In *IJCAI*, pages 4623–4629, 2018.

[Zhu *et al.*, 2017] Wenya Zhu, Kaixiang Mo, Yu Zhang, Zhangbin Zhu, Xuezheng Peng, and Qiang Yang. Flexible end-to-end dialogue system for knowledge grounded conversation. *CoRR*, abs/1709.04264, 2017.