# Diverse and Informative Dialogue Generation with Context-Specific Commonsense Knowledge Awareness

**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, 100871, China
[2]National Research Center of Software Engineering,
Peking University, Beijing, 100871, China
[3]Auburn University, Auburn, Alabama, 36849, USA
{wusixing, li.ying, daweizhang}@pku.edu.cn
yangzhou@auburn.edu, zhwu@ss.pku.edu.cn

## Abstract

Generative dialogue systems tend to produce generic responses, which often leads to boring conversations. For alleviating this issue, Recent studies proposed to retrieve and introduce knowledge facts from knowledge graphs. While this paradigm works to a certain extent, it usually retrieves knowledge facts only based on the entity word itself, without considering the specific dialogue context. Thus, the introduction of the context-irrelevant knowledge facts can impact the quality of generations. To this end, this paper proposes a novel commonsense knowledge-aware dialogue generation model, ConKADI. We design a Felicitous Fact mechanism to help the model focus on the knowledge facts that are highly relevant to the context; furthermore, two techniques, Context-Knowledge Fusion and Flexible Mode Fusion are proposed to facilitate the integration of the knowledge in the ConKADI. We collect and build a large-scale Chinese dataset aligned with the commonsense knowledge for dialogue generation. Extensive evaluations over both an open-released English dataset and our Chinese dataset demonstrate that our approach ConKADI outperforms the state-of-the-art approach CCM, in most experiments.

## 1 Introduction

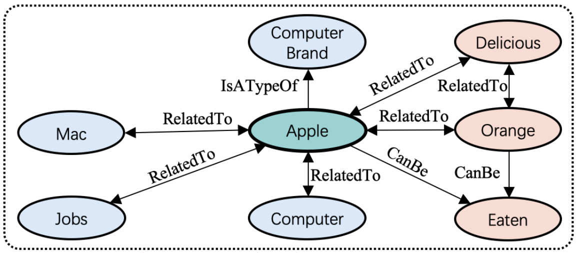
Nowadays, open-domain dialogue response generation systems have shown impressive potential, to endow a machine with the ability to converse with a human, using natural language (Chen et al., 2017). Although such models have achieved promising performance, they still suffer from generating generic and boring responses, such as ”I don’t know.” Such low-quality responses always reduce the attractiveness of generative dialogue systems to end-users. Researchers have tried to tackle it from multiple aspects; for example, using the enhanced objective function (Li et al., 2016a); introducing additional contents (Xu et al., 2019). However, these methods haven’t solved the issue thoroughly. Different from a human being, who is capable of associating the dialogue with the background knowledge in his/her mind, a machine can merely capture limited information from the surface text of the query message (Ghazvininejad et al., 2018). Consequently, it is difficult for a machine to understand the query fully, and then to generate diverse and informative responses (Zhou et al., 2018).

To bridge the gap of the knowledge between the human and the machine, researchers have begun to introduce large-scale knowledge graphs for enhancing the dialogue generation (Zhu et al., 2017; Zhou et al., 2018; Liu et al., 2018), and they have obtained lots of impressive results. Generally, the retrieval of knowledge facts is based on the entity name; in detail, the first step is to recognize entity words in the given query message, and then facts that contain the mentioned entities can be retrieved as candidates[1]. Subsequently, a knowledge-aware response can be generated based on the query message and previously retrieved facts. Although such a straightforward paradigm works to a certain extent, some challenges in knowledge-aware dialogue generation still keep unsolved. **1)** An entity word usually can refer to different concepts, i.e., an entity has multiple meanings, but only one specific concept is involved in a particular context. Without considering this, some pre-fetched knowledge fact candidates can be irrelevant to the context. **2)** Even if we only consider a particular entity meaning, the related knowledge facts may cover various target topics. However, some of those topics do not con-

[*] Corresponding author: Ying Li, li.ying@pku.edu.cn

[1] For example, for a mentioned entity “apple” in a query, the fact (apple, is a type of, fruit) or (fruit, related to, apple) can be retrieved.

*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5811–5820
July 5 - 10, 2020. ©2020 Association for Computational Linguistics



**Message**: Apple's new product is awesome!

**#1**: Yes, a beautiful new **Mac**.

**#2**: I love it, as delicious as the orange.

Figure 1: An illustrative example. #1 shows the response generated with a highly relevant fact, #2 shows the response generated with irrelevant facts.

tribute to the dialogue generation. Figure 1 presents an illustrative example to demonstrate such two issues. Here, a subgraph is retrieved based on the entity word "Apple" in the query. In general, "Apple" can be interpreted as either a type of fruit or a brand name. In this context, it is evident that "Apple" refers to a brand name. However, some knowledge facts concerning a type of fruit are retrieved too. If a model makes an inappropriate choice of irrelevant facts, the generated response will make no sense to the query message. In our example, even for the entities in blue circle related to the brand name "Apple", only some of them have a positive effect in the dialogue generation, e.g., "Jobs" should not make any contribution to the "#1". **3)** The integration of the knowledge and the dialogue generation in previous approaches is insufficient, including the way of integration, as well as the types of knowledge.

To tackle such challenges, this paper proposes a Context Knowledge-Aware Diverse and Informative conversation generation model, ConKADI. **First**, we design a Felicitous Fact mechanism to help the model highlight the knowledge facts that are highly relevant to the context, that is, "Felicitous Facts". Felicitous Fact mechanism generates a felicitous fact probability distribution over the retrieved facts. For improving the selection of felicitous facts, human-generated answers (i.e., the ground-truth responses) are used as the posterior context knowledge to supervise the training of the prior felicitous fact probability distribution. **Next**, Context-Knowledge Fusion is proposed to lift the role of knowledge facts in the dialogue generation, by fusing the context and the felicitous knowledge before the decoding. **Last**, ConKADI can generate three types of words owing to the Flexible Mode Fusion module, which aims at simultaneously fusing multiple types of knowledge. To summarize, Felicitous Fact mechanism can alleviate the first two issues, and the next two techniques solve the last issue. Consequently, our approach can improve the utilization rate of knowledge graphs, as well as can promote the diversity and informativeness of the generated responses.

In the experiments, a large-scale Chinese Weibo dataset is collected and aligned with the commonsense knowledge for dialogue generation. We perform extensive evaluations on two large-scale datasets: an open-released English Reddit dataset and our proposed Chinese Weibo dataset. The experimental results demonstrate that our proposed ConKADI model significantly outperforms representative methods in knowledge utilization, diversity, and informativeness. Especially, ConKADI exceeds the latest knowledge-aware dialogue generation model, CCM (Zhou et al., 2018), in most experiments.

## 2 Related Work

Seq2Seq (Sutskever et al., 2014; Vinyals and Le, 2015) has been widely used in the open-domain dialogue generation. However, models tend to generate generic responses (Serban et al., 2016). To tackle this issue, researchers have proposed new objectives (Li et al., 2016a), enhanced decoding algorithms (Li et al., 2016b), latent-variable based methods (Zhao et al., 2017, 2018; Gao et al., 2019). Introducing additional contents into the dialogue generation is also helpful. (Xu et al., 2019) uses meta-words; (Zhu et al., 2019) uses the retrieved existing dialogues. However, the leading cause of generating generic responses is that the model can not obtain enough background knowledge from the query message (Ghazvininejad et al., 2018; Liu et al., 2019).

Recently, to alleviate the lack of background knowledge, researchers have begun to introduce the knowledge into the generation. The knowledge can be the unstructured knowledge texts (Ghazvininejad et al., 2018), the structured knowledge graphs (Zhou et al., 2018), or the hybrid of them (Liu et al., 2019). The structured knowledge has the best quality, because it is generally extracted and summarized by the human. The structured knowledge graph can be either domain-specific (Zhu et al., 2017; Liu et al., 2018) or open-domain (Young et al., 2018; Zhou et al., 2018). ConceptNet (Speer et al., 2017) is a multilingual open-domain commonsense knowledge graph, which is designed to represent the general knowledge and to improve understanding of the meanings behind the words people use. Two previous studies (Young et al., 2018; Zhou et al., 2018) have proved the feasibility of introducing commonsense knowledge into dialogue systems. The first work (Young et al., 2018) is designed for retrieval-based systems; therefore, only the current state-of-the-art CCM (Zhou et al., 2018) is our direct competitor. In comparison with CCM, 1) ConKADI is aware of the context when using the knowledge. 2) ConKADI uses human's responses as posterior knowledge in training.

In addition, our Felicitous Fact mechanism is different from the word/knowledge selection mechanisms previously proposed in related tasks; for example, selecting a cue word (Mou et al., 2016; Yao et al., 2017) or selecting a knowledge (Liu et al., 2019). First, ConKADI can access more contextual information because our model is fully end-to-end, while previous works use independent and external modules. Second, our Felicitous Fact outputs a probabilistic distribution instead of a hard singleton value, as did the previous works.

## 3 Approach

### 3.1 Task Formulation and Model Overview

Formally, given a training data $\mathcal{D}$ of triplets, where each triplet includes a query message $X = (x_1, \ldots, x_n)$, a response $Y = (y_1, \ldots, y_m)$, and a set of commonsense knowledge facts $F = \{f_1, \ldots, f_l\}$. The training goal of knowledge-aware dialogue generation is to maximize the probability $\sum_{(X,Y,F)\in\mathcal{D}} \frac{1}{|\mathcal{D}|} p(Y|X,F)$; the inference goal is to find $Y^* = \arg\max_Y p(Y|X,F)$. Knowledge facts $F$ are retrieved from the knowledge graph $\mathcal{G}$ ; each fact is organized as a triplet $(h, r, t)$.

The overview of ConKADI has been shown in Figure 2. Knowledge fact set $F$ is retrieved by the **Knowledge Retriever** given the query message $X$. The **Context Encoder** summarizes an utterance into contextual representations. The **Felicitous Fact Recognizer** calculates the felicitous fact probability distribution $\mathbf{z}$ over the $F$, which is used to initialize the Decoder and guide the generation. The **Triple Knowledge Decoder** can generate three types of words: vocabulary words, entity words, and copied words, with the **Flexible Mode Fusion**.

### 3.2 Felicitous Fact mechanism

**Knowledge Retriever:** Given a query message $X$, if a word $x_i \in X$ is recognized as an entity word and can be matched to a vertex $e_{src}$ in the knowledge graph $\mathcal{G}$, then, each neighbour $e_{tgt} \in Neighbour(e_{src})$ and the corresponding directional relation $r$ is retrieved as a candidate fact $f$. $e_{src}$/$e_{tgt}$ is called as **source/target** entity. If a word can't match any vertex, a special fact $f_{NAF}$ will be used.

**Context Encoder:** The Context Encoder is a bi-directional GRU network (Cho et al., 2014), which reads $X$ or $Y$ and outputs a contextual state sequence. For simplicity, we take $X$ as an example. At the time step $t$, the Encoder outputs a forward state and a backward state, the concatenation of such two states $\mathbf{h_t^x} = [\mathbf{h_t^{fw}}; \mathbf{h_t^{bw}}] \in \mathbb{R}^{2d_h \times 1}$ is regarded as the contextual state :

$$\begin{aligned} \mathbf{h_t^{fw}} &= GRU^{fw}(\mathbf{h_{t-1}^{fw}}, \mathbf{x_t}, \mathbf{e_{x_t}}) \\ \mathbf{h_t^{bw}} &= GRU^{bw}(\mathbf{h_{t-1}^{bw}}, \mathbf{x_{n-t+1}}, \mathbf{e_{x_{n-t+1}}}) \end{aligned} \tag{1}$$

where $\mathbf{x_t}$ is the word embedding of $x_t$. To enhance the semantic information, the matched entity embedding $\mathbf{e_{x_t}}$ of $x_t$ is also involved. Finally, the contextual state sequence of $X/Y$ is denoted as $H^{x/y} = (h_1^{x/y}, \ldots, h_{n/m}^{x/y})$. Specifically, $H^x$ is the prior context; $H^y$ is the posterior context that is only available in the training stage.

**Felicitous Fact Recognizer:** Recall the example illustrated in Figure 1 , some preliminary retrieved knowledge facts may be inappropriate in the dialogue context. The Felicitous Fact Recognizer is designed to detect the facts that highly coincide with the dialogue context, i.e., Felicitous Facts. The Felicitous Fact Recognizer reads the contextual information, then outputs a probability distribution $\mathbf{z} \in \mathbb{R}^{l \times 1}$ over the $F$; therefore, the $i$-th dimension value $\mathbf{z}[i]$ indicates the weight of $f_i$. In the training stage, the high-quality human-generated response $Y$ is served as the posterior knowledge; hence, the posterior $\mathbf{z_{post}}$ is adopted in training, the prior $\mathbf{z_{prior}}$ is adopted in inference:

$$\mathbf{z_{post}} = \eta(\varphi(\mathbf{F} \cdot \mathbf{W_{ft}}) \cdot \varphi([\mathbf{h_n^x}^\top; \mathbf{h_m^y}^\top] \cdot \mathbf{W_{post}}))^\top$$

$$\mathbf{z_{prior}} = \eta(\varphi(\mathbf{F} \cdot \mathbf{W_{ft}}) \cdot \varphi(\mathbf{h_n^x}^\top \cdot \mathbf{W_{prior}}))^\top \tag{2}$$

where $\mathbf{F} \in \mathbb{R}^{l \times (d_e + d_r + d_e)}$ is the embedding matrix of the retrieved facts $F$, $\mathbf{W_{ft}}$, $\mathbf{W_{post}}$ and $\mathbf{W_{prior}}$ are trainable parameters, $\eta$ is $softmax$ activation ,

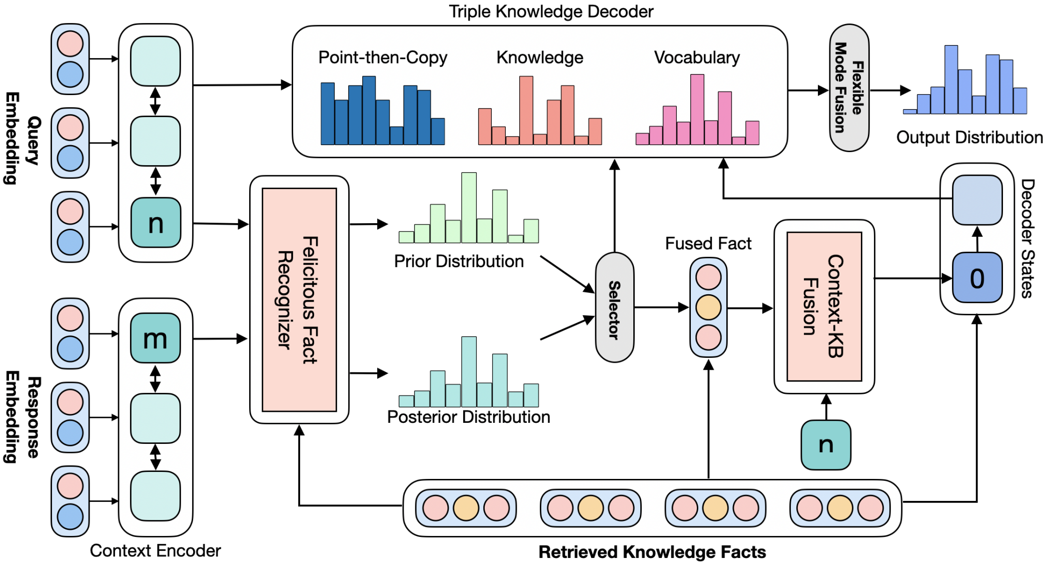


Figure 2: An overview of the proposed approach ConKADI.

$\varphi$ is $tanh$ activation. Kullback–Leibler Divergence (Kullback and Leibler, 1951) (**KLD**) is used to force two distributions to become as close as possible.

$$\mathcal{L}_k = \mathbf{KLD}(\mathbf{z_{post}}, \mathbf{z_{prior}}) \qquad (3)$$

**Context-Knowledge Fusion:** To enhance the Decoder's understanding of the background knowledge, the Decoder is initialized based on the fused knowledge $\mathbf{f_z}^\top = \mathbf{z}^\top \cdot \mathbf{F}$ and the query context:

$$\mathbf{h_0^y}^\top = tanh([\mathbf{h_n^x}^\top; \mathbf{f_z}^\top] \cdot \mathbf{W_{init}}) \qquad (4)$$

where $\mathbf{W_{init}}$ is a trainable parameter.

Following the previous work (Zhao et al., 2017), we adopt the Bag-of-Words Loss to ensure the accuracy of the input of the Context-Knowledge Fusion, namely, $\mathbf{h_n^x}$ and $\mathbf{f_z}$. Meanwhile, we construct a 0-1 indicator vector $\mathbf{I_f} \in \mathbb{R}^{l\times 1}$ to supervise the training of $\mathbf{z_{post}}$, where $\mathbf{I_f}[i]$ is set to 1 if the target entity of the $i$-th fact $f_i$ appears in the $Y$, otherwise 0. Thus, the objective is to minimize the $\mathcal{L}_f$ given by:

$$-\frac{\sum_{y_b \in B} \log p_b(y_b|\mathbf{h_n^x}, \mathbf{f_z}))}{|B|} - \frac{\mathbf{I_f}^\top \cdot \log(\mathbf{z_{post}})}{|\mathbf{I_f}|} \qquad (5)$$

where B is the word bag of $Y$, $p_b$ is a 2-layer $MLP_{bow}$ activated by $softmax$, which outputs the probability distribution over the vocabulary $V$.

### 3.3 Triple Knowledge Decoder

The Decoder is another GRU network. At each time step, the Decoder can generate one of three types of words: vocabulary words, knowledgeable entity words, and copied words. ConKADI first updates the internal state:

$$\mathbf{h_t^y} = g(\mathbf{h_{t-1}^y}, \mathbf{u_{t-1}}, \mathbf{c_{t-1}}) \qquad (6)$$

where $\mathbf{u_{t-1}}^\top = [\mathbf{y_{t-1}^\top}; \mathbf{e_{y_{t-1}}^\top}; \mathbf{h_{y_{t-1}}^x}^\top]$, and $\mathbf{y_{t-1}}$, $\mathbf{e_{y_{t-1}}}$, $\mathbf{h_{y_{t-1}}^x}$ are the word embedding, the entity embedding and the pointed-then-copied source state of the last predicted token $y_{t-1}$, respectively; and $\mathbf{c_{t-1}}$ is the Attention [2].

**Vocabulary Words:** The probability distribution $p_{w,t} \in \mathbb{R}^{|V|\times 1}$ over the $V$ is given by:

$$p_{w,t}^\top = \eta(elu([\mathbf{h_t^y}^\top; \mathbf{u_{t-1}}^\top; \mathbf{c_t}^\top] \cdot \mathbf{W_{v1}}) \cdot \mathbf{W_{v2}}) \qquad (7)$$

where $\mathbf{W_{v1/2}}$ are trainable parameters, and the non-linear activation $elu$ is proposed by (Clevert et al., 2016).

**Knowledgeable Entity Words:** An entity word can be generated by extracting the target entity of the best-matched fact $f$ at each time step. The corresponding probability distribution $p_{k,t} \in \mathbb{R}^{l\times 1}$ over the $F$ is calculated as:

$$\mathbf{z_{d,t}} = \eta(\varphi(\mathbf{F} \cdot \mathbf{W_{fd}}) \cdot \varphi([\mathbf{h_t^y}^\top; \mathbf{u_{t-1}}^\top] \cdot \mathbf{W_d})^\top)$$

$$\gamma_t = sigmoid([\mathbf{h_t^y}^\top; \mathbf{u_t}^\top; \mathbf{c_t}^\top] \cdot \mathbf{W_{gate}}) \in \mathbb{R}^1$$

$$p_{k,t} = \gamma_t \times \mathbf{z} + (1.0 - \gamma_t) \times \mathbf{z_d} \qquad (8)$$

[2] We have omitted the description of Attention. Please see (Luong et al., 2015) for the detail.

where the previous $\mathbf{z}$ here serves as a static global distribution (denoted as **GlFact**), $\mathbf{z_{d,t}}$ is the dynamic distribution, and $\gamma_t$ is a gate to control the contribution of each distribution.

**Copied Words:** The Decoder can further point out a word $x$ from $X$, and then copies the $x$ . The corresponding probability distribution $p_{c,t} \in \mathbb{R}^{n\times 1}$ over the query message $X$ is calculated as:

$$\begin{aligned} p_{c,t} &= \eta(\varphi(\mathbf{H^x}\cdot\mathbf{W_{cs}})\cdot\varphi(\mathbf{u_t^{c\top}}\cdot\mathbf{W_{ct}})^\top) \\ \mathbf{u_t^{c\top}} &= [\mathbf{h_t^{y\top}};\mathbf{u_{t-1}}^\top;\mathbf{c_t}^\top] \end{aligned} \tag{9}$$

**Flexible Mode Fusion:** Previous three distributions can be fused by the $MF(\mathbf{h_t^y},\mathbf{u_{t-1}},\mathbf{c_t})$, a 2-layer MLP activated by $softmax$. $MF$ can outputs a probability distribution $(\gamma_{w,t},\gamma_{k,t},\gamma_{c,t})$ over three modes at each time step:

$$p_{out,t} = \gamma_{w,t}\times p_{w,t} + \gamma_{k,t}\times p_{k,t} + \gamma_{c,t}\times p_{c,t} \tag{10}$$

The proposed $MF$ can be regarded as a multi-class classifier; therefore, the advantage of $MF$ is the flexibility, we can additionally integrate more modes or remove existing modes by simply changing the number of classes. For a more reasonable fusion, the Cross-Entropy between the ground-truth mode and the predicted distribution by $MF$ is used to supervise the training; the corresponding Cross-Entropy loss is denoted as $\mathcal{L}_m$. Next, we optimize the fused output distribution $p_{out}(Y|X,F)$ by minimizing the $\mathcal{L}_n$, which is given by:

$$-\sum_t \lambda_t \log p_{out,t}(y_t|y_{t-1:1},X,F) + \frac{\mathcal{L}_m}{2} \tag{11}$$

where $\lambda_t$ is a normalization term to penalize the out-of-vocabulary words, $\lambda_t = \frac{1}{\#(unk\in Y)}$ [3] if $y_t$ is an $unk$, otherwise $\lambda_t = 1$.

**Training Objective:** Finally, the ConKADI can be trained by minimizing the following objective:

$$\mathcal{L} = \mathcal{L}_n + \mathcal{L}_k + \mathcal{L}_f \tag{12}$$

# 4 Experiments

## 4.1 Dataset

To verify the generalization among different languages, we evaluate models not only on a public English Reddit dataset (Zhou et al., 2018), but we also collect and construct a Chinese Weibo dataset. Both datasets are aligned with the commonsense knowledge graph ConcetNet (conceptnet.io), the statistics have been reported in Table 1.

| | Reddit | Weibo |
|---|---|---|
| #Train | 1,352,961 | 1,019,908 |
| #Dev/#Test | 40,000 | 56,661 |
| #Vocab | 30,000 | 50,000 |
| Batch Size | 100 | 50 |
| #Entity/#Relation | 21,471/44 | 27,189/26 |
| #Fact | 149,803 | 696,466 |

Table 1: The statistics of two datasets.

**The English Reddit:** We did some filtering on the raw data: Utterances that are too short (< 4 words) or too long (> 30 words) were removed, and each message can be associated with at most 300 related fact triplets.

**The Chinese Weibo:** We first collected three open-sourced Weibo (weibo.com) datasets, which originally contained 4.44M (Shang et al., 2015), 1.96M (Ke et al., 2018) and 10.48M (Li and Yan, 2018) pairs of dialogue, respectively. Jieba[4] was used to segment; utterances that are too short/long were removed as well. Next, we crawled 4.48M entities and 13.98M facts from the ConceptNet. Stop entities, and low-frequent entities are excluded. For a dialogue pair, if one entity in the message and another entity in the response can be connected by a 1-hop edge in the knowledge graph, this dialogue was kept. In comparison with the English Reddit, our dataset has more facts, but the relation types are quite limited; hence, we set the limit that a message can be associated with at most 150 fact triplets.

For two datasets, the embedding of entities and relations are learned by using TransE (Bordes et al., 2013); then, they are kept fixed in training. Our experimental resources are available at the web [5].

## 4.2 Settings

**Baselines:** The widely used **S2S** (Sutskever et al., 2014), and its Attentive version **ATS2S** (Luong et al., 2015). We further add the bidi-MMI (Li et al., 2016a) or the diverse decoding (Li et al., 2016b) to improve the diversity of ATS2S, which are denoted as **ATS2S**$_{MMI}$ and **ATS2S**$_{DD}$[6]. **Copy** mechanism (Gu et al., 2016; Vinyals et al., 2015) allows Decoder to point then copy a source word. **GenDS** is a knowledge-aware model, which can generate responses with the utilizing of entity words. (Zhu et al., 2017). **CCM** is the current state-of-the-art approach in the task of response generation with

[3] #(·) is the count of ·

[4] https://pypi.python.org/pypi/jieba/

[5] https://github.com/pku-orangecat/ACL2020-ConKADI

[6] The best $k$ was searched form $[0.1, 3.0]$.

|  | Entity Score |  |  | Embedding |  | Overlap (%) |  | Diversity (%) |  | Informativeness | R-Score |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Metric | E$_{match}$ | E$_{use}$ | E$_{recall}$ | Emb$_{avg}$ | Emb$_{ex}$ | BLEU-2 | BLEU-3 | Distinct-1 | Distinct-2 | Entropy | R$_a$ | R$_g$ |
| **Chinese Weibo** |  |  |  |  |  |  |  |  |  |  |  |  |
| S2S | 0.33 | 0.58 | 13% | 0.770 | 0.500 | 2.24 | 0.80 | 0.21 | 1.04 | 6.09 | 0.78 | 0.75 |
| ATS2S | 0.33 | 0.59 | 12% | 0.767 | 0.513 | 1.93 | 0.69 | 0.27 | 1.23 | 5.99 | 0.77 | 0.75 |
| ATS2S$_{MMI}$ | 0.40 | 0.74 | 15% | 0.773 | 0.528 | 4.01 | **1.61** | 0.75 | 3.91 | 7.49 | 1.24 | 1.21 |
| ATS2S$_{DD_{1.5}}$ | 0.35 | 0.62 | 13% | 0.780 | 0.542 | 2.14 | 0.86 | 1.03 | 4.86 | 7.62 | 1.16 | 1.10 |
| Copy | 0.33 | 0.68 | 13% | 0.786 | 0.501 | 2.28 | 0.84 | 0.59 | 2.18 | 6.13 | 0.92 | 0.91 |
| GenDS | 0.75 | 0.84 | 26% | 0.789 | 0.524 | 2.09 | 0.73 | 0.30 | 1.66 | 5.89 | 0.94 | 0.91 |
| CCM | 0.99 | 1.09 | 28% | 0.786 | 0.544 | 3.26 | 1.20 | 0.48 | 2.59 | 6.16 | 1.18 | 1.15 |
| AVG | 0.49 | 0.74 | 17% | 0.779 | 0.522 | 2.56 | 0.96 | 0.52 | 2.50 | 6.48 | 1.00 | 1.00 |
| ConKADI | 1.48 | **2.08** | **38%** | **0.846** | **0.577** | **5.06** | 1.59 | **3.26** | **23.93** | **9.04** | **2.98** | **2.24** |
| ConKADI$_{-cp}$ | **1.60** | 1.89 | **38%** | 0.833 | 0.567 | 5.00 | 1.52 | 2.34 | 18.29 | 8.75 | 2.55 | 2.08 |
| **English Reddit** |  |  |  |  |  |  |  |  |  |  |  |  |
| S2S | 0.41 | 0.52 | 4% | 0.868 | 0.837 | 4.81 | 1.89 | 0.38 | 1.77 | 7.59 | 0.82 | 0.78 |
| ATS2S | 0.44 | 0.59 | 5% | 0.863 | 0.831 | 4.50 | 1.81 | 0.82 | 3.44 | 7.62 | 0.92 | 0.91 |
| ATS2S$_{MMI}$ | 0.45 | 0.65 | 6% | 0.858 | 0.825 | 4.95 | 2.13 | 0.75 | 3.22 | 7.62 | 0.95 | 0.94 |
| ATS2S$_{DD_{0.3}}$ | 0.31 | 0.43 | 4% | 0.830 | 0.784 | 1.70 | 0.75 | 0.97 | 3.50 | 7.47 | 0.77 | 0.72 |
| Copy | 0.13 | 0.67 | 9% | 0.868 | 0.841 | **5.43** | **2.26** | 1.73 | 8.33 | 7.87 | 1.19 | 1.09 |
| GenDS | 1.13 | 1.26 | 13% | **0.876** | 0.851 | 4.68 | 1.79 | 0.74 | 3.97 | 7.73 | 1.14 | 1.10 |
| CCM | 1.08 | 1.33 | 11% | 0.871 | 0.841 | 5.18 | 2.01 | 1.05 | 5.29 | 7.73 | 1.21 | 1.18 |
| AVG | 0.55 | 0.77 | 7% | 0.860 | 0.829 | 4.40 | 1.79 | 0.94 | 4.32 | 7.69 | 1.00 | 1.00 |
| ConKADI | 1.24 | **1.98** | **14%** | 0.867 | 0.852 | 3.53 | 1.27 | **2.77** | **18.78** | 8.50 | **1.76** | **1.46** |
| ConKADI$_{-cp}$ | **1.41** | 1.73 | 13% | 0.865 | **0.855** | 3.09 | 1.07 | 2.29 | 16.70 | **8.68** | 1.63 | 1.37 |

Table 2: Objective Experimental Results. The ablation ConKADI$_{-cp}$ removes the ability to copy source words.

commonsense knowledge (Zhou et al., 2018).

**Implementation:** We implemented all models except CCM, CCM was tested based on its official code[7]. Most hyper-parameters are kept the same as CCM, and hyper-parameters among models are kept the same as possible. In detail, the word embedding dimension is 300, Encoder is a 2-layer bidirectional GRU with 512 units, and Decoder is a 2-layer GRU with 512 units. Adam is used to optimizing model with an initial learning rate $lr = 0.0001$; if $perplexity$ begins to increase, the $lr$ will be halved, if $perplexity$ increases in two continuous epochs, the training will be stopped. Following the CCM, the maximum epoch number is 20.

**Objective Metrics:** We evaluate the generated responses from four aspects: **Knowledge Utilization ($\mathbf{A_1}$)**: E$_{match}$ is the averaged number of the matched target entities per generation. (Zhou et al., 2018). E$_{use}$ further counts the source entities. E$_{recall}$ is the ratio of recalled entities. **Embedding-based Relevance ($\mathbf{A_{2a}}$)** : Following (Liu et al., 2016), we use the Emb$_{avg}$ that considers the averaged word embedding, and the Emb$_{ex}$ that considers each dimension's extreme value. **Overlapping-based Relevance ($\mathbf{A_{2b}}$)** : BLEU-2/3 (Tian et al., 2017; Wu et al., 2017). **Diversity ($\mathbf{A_3}$)**: We report the ratio of distinct uni/bi-grams, i.e., Distinct-1/2, in all generated texts (Li et al., 2016a; Wu et al., 2018). **Informativeness ($\mathbf{A_4}$)**: We report the word-level Entropy (Mou et al., 2016).

**Relative Score:** To illustrate the comprehensive performance of models, we first compute the average score of 7 baselines metric by metric (**AVG**), then, we report the arithmetic mean score:

$$R_a = \frac{1}{5}\sum_{A_i}\left(\frac{1}{|A_i|}\sum_{m \in A_i}\frac{m_j}{m_{j,AVG}}\right) \tag{13}$$

and the geometric mean score:

$$R_g = \left(\prod_{A_i}\left(\prod_{m_j \in A_i}\frac{m_j}{m_{j,AVG}}\right)^{\frac{1}{|A_i|}}\right)^{\frac{1}{5}} \tag{14}$$

### 4.3 Experimental Results

The objective evaluation results on the two datasets have been reported in Table 2. By reviewing the **Relative Score**, it can be seen that the overall performance of ConKADI outperforms baseline models. More specifically, our ConKADI outperforms baseline models in terms of all metrics except BLEU-3 on the Chinese Weibo, and our ConKADI outperforms baseline models in terms of almost all metrics on the English Reddit. In comparison with the state-of-the-art method CCM, our ConKADI increases the overall performance by 153%/95% (arithmetic/geometric mean) on the Chinese dataset, as well as increases the overall performance by 48%/25% on the English dataset.

[7] CCM doesn't support beam-search, so we use the greedy search except ATS2S$_{MMI}$ and ATS2S$_{DD}$ use beam=10.

**Knowledge Utilization:** By accessing the knowledge, three knowledge-aware models, i.e., GenDS, CCM, and ConKADI, can significantly outperform other models. In comparison with GenDS and CCM, the advantages of ConKADI can be summarized as 1) ConKADI has a higher utilization of the knowledge, which can be proved by E$_{match}$. 2) By using the point-then-copy mechanism (ConKADI vs. ConKADI$_{-cp}$), ConKADI further expands the total generated entity number (E$_{use}$). After adding the point-then-copy mechanism, while the E$_{match}$ drops by 7.5%, the overall E$_{use}$ increases by 10%. It means ConKADI can reasonably decide whether to use a knowledge fact or copy a source word. 3) ConKADI is more potential to find out the accurate knowledge; hence, our E$_{recall}$ is much higher than the E$_{recall}$ of GenDS and CCM. Such results can demonstrate that the proposed Felicitous Fact mechanism can help the model better focus on the facts that are relevant to the dialogue context, and increase the utilization rate of the knowledge graph and the accuracy of the knowledge selection.

**Diversity and Informativeness:** Generative models have been suffering from generating responses without enough diversity and informativeness. Although previous GenDS and CCM can utilize the knowledge, they fail to solve this challenge; they even can be beaten by other baselines. By contrast, our ConKADI has significantly alleviated this issue. According to our ablation experiments, such notable promotion can be attributed to the proposed Context-Knowledge Fusion. The more detail will be discussed in the ablation study.

**Relevance:** On the Chinese dataset, ConKADI has the best overall performance, but ConKADI's performance is not ideal on the English dataset. First, we think the reason is the inherent difference of datasets; two datasets are collected from different sources and have varying densities of entity-relations (see Table 1). Next, we must emphasize these metrics can only evaluate the relevance to the given reference. Instead of the 1-to-1 mapping, the dialogue is undoubtedly a 1-to-n mapping; therefore, these results cannot show the generation is not consistent with the query. ConKADI is a very diverse model; only use one reference to judge is unfair. Similarly, this limitation has been found and explained in a recent work (Gao et al., 2019).

### 4.4 Human Annotation

| ConKADI | Appropriateness | | | Informativeness | | |
|---|---|---|---|---|---|---|
| vs. | Win | Tie | Lose | Win | Tie | Lose |
| ATS2S | 71.3% | 11.0% | 17.7 % | 87.3% | 6.9% | 5.8% |
| ATS2S$_{MMI}$ | 59.3% | 9.2% | 31.5% | 82.5% | 7.3% | 10.2% |
| Copy | 71.7% | 8.8% | 19.5% | 89.7% | 3.8% | 6.5% |
| GenDS | 87.2% | 7.3% | 5.5% | 93.8% | 2.3% | 3.5% |
| CCM | 83.8% | 6.9% | 9.3% | 93.0% | 3.5% | 3.5% |

Table 3: Human annotation results on the Chinese Weibo. ConKADI significantly (sign test, p-value < 0.005, ties are removed) outperforms other baselines in terms of both appropriateness and informativeness.

Following (Liu et al., 2019), we randomly sample 200 query messages from the test set, and then we conduct the pair-wise comparison. For the variations of S2S, We remain two most representative models, ATS2S and ATS2S$_{MMI}$. Thus, we have 1,000 pairs in total. For each pair, we invite three well-educated volunteers to judge which response is better, in terms of the following two metrics: 1) Appropriateness, which mainly considers the fluency and the logical relevance. 2) Informativeness, which considers whether the model provides new information/knowledge or not. The tie is allowed, but volunteers are required to avoid it as possible. The model names are masked, and the A-B order is random.

For the appropriateness, 2/3 agreement (i.e., the percentage of cases that at least 2 volunteers give the same label) is 95%, and the 3/3 agreement is 67.1%. For the informativeness, 2/3 agreement is 97%, and the 3/3 agreement is 79.1%.

The results have been reported in Table 3. ATS2S$_{MMI}$ is the strongest baseline owing to the beam search and the MMI re-ranking, especially in terms of appropriateness. While the generation of ATS2S$_{MMI}$ is more generic, it's friendly for human reading; hence, it tends to receive higher scores. GenDS and CCM are far behind our model. We find their generation is usually not fluent, while a lot of entities are generated. Comparing two metrics, ConKADI has more notable advantages in terms of informativeness.

### 4.5 Ablation Study

We focus on the ablation of the Felicitous Fact mechanism. There are 3 factors, GlFact (using the distribution **z** to guide the entity word generation), CKF (Context-Knowledge Fusion), and CKF's loss $\mathcal{L}_f$. **Copy** has fully removed the Felicitous Fact mechanism (i.e., above 3 factors); **Base** further

| Query | #1:My cat likes bananas and bread. | #2:Yeah , but what website? | #3:我会唱霉霉的歌。<br>I can sing the song of Taylor Swift |
|---|---|---|---|
| ATS2S | I'm a banana and I don't know what you're talking about. | I'm not sure. I'm just curious. | 我也是，我唱的是 **unk**。<br>Me too. I'm singing **unk**。 |
| ATS2S$_{MMI}$ | Do you have a cat? | It's a site site. | 你唱的是哪种歌?<br>What kind of song are you singing? |
| Copy | I'm a cat. | I'm not sure what site you're talking about. | 我也是，我也是，我也是，我也喜欢。<br>Me too, me too, me too, I like it。 |
| GenDS | I'm a banana. | I'm not sure , but I'm not sure if it's a link to the original post. I'm not sure what the site is. | 你可以听我唱的唱。<br>You can listen to my singing singing. |
| CCM | I'm a banana and I love my cat. | I'm not sure, I just got a link to the site. | 我也是,我也喜欢,听着歌着歌听着歌听着歌<br>Me too. I like it, too. Listening to songs. Listening to songs. Listening to songs |
| ConKADI | And your cat is the best. | Looks like Youtube, the site is blocked. | 我听了,他的音乐好听。<br>I heard it. His music is good. |

Table 4: Case Study: #1 #2 are sampled from the English Reddit, #3 is sampled from the Chinese Weibo.

removes the ability to copy source words.

| # | Settings | E$_{use}$ | Distinct-2 | Entropy | R$_g$ |
|---|---|---|---|---|---|
| #1 | Copy+GlFact+CKF+$\mathcal{L}_f$ | 2.08 | 23.93 | 9.04 | 2.24 |
| #2 | Base+GlFact+CKF+$\mathcal{L}_f$ | 1.89 | 18.29 | 8.75 | 2.02 |
| #3 | Copy+GlFact+CKF | 1.79 | 18.18 | 8.73 | 2.08 |
| #4 | Base+GlFact+CKF | 1.92 | 17.38 | 8.87 | 2.01 |
| #5 | Base+CKF | 1.87 | 15.72 | 8.66 | 1.96 |
| #6 | Base+GlFact | 1.05 | 2.90 | 6.31 | 1.10 |
| #7 | Base | 1.06 | 2.50 | 6.46 | 1.10 |

Table 5: Ablation study on the Chinese Weibo.

The results have been reported in Table 5. 1) The performance drops significantly without using the context-knowledge fused result to initialize the Decoder (#5 → #7), indicating that CKF is very important for the Decoder. 2) If GlFact is adopted solely, it can affect performance in turn. 3) $\mathcal{L}_f$ is essential to the Copy in comparison with Base.

**Analysis of KL Divergence:** The training stage introduces posterior knowledge, which is absent during the inference. Therefore, reducing the difference between such two distribution is very necessary. We here check the curve of the **KLD** between the $\mathbf{z_{prior}}$ and $\mathbf{z_{post}}$, i.e., $\mathcal{L}_k$ . A lower $\mathcal{L}_k$ means the two distribution are closer. As shown in Figure 3: 1) **KLD** is strongly related to the overall performance. 2) The importance that using the fused knowledge to initialize the Decoder (CKF) has been proved once again (#5 vs. #6).

### 4.6 Case Study

Three cases are sampled in Table 4. In case 1, except ATS2S$_{MMI}$ and our ConKADI, the remaining models have generated weird responses. ATS2S$_{MMI}$ generated a fluent response, but this response is not very logically relevant to the query. In case 2, although GenDS and CCM have generated entity words, they also generate some redundant generic patterns, namely, "I'm not sure ...". It is perhaps because their understanding of background knowledge is still not enough. Our ConKADI generates a fluent and informative response. The last challenging case is sampled from the Chinese dataset. "Taylor Swift" is a female singer, but it is an unknown word for models. All generated responses are not absolutely perfect. Only the generations of ATS2S$_{MMI}$ and ConKADI are fluent. In comparison with ATS2S$_{MMI}$, the generation of ConKADI provides more information; the only small flaw is ConKADI wrongly thinks "Taylor Swift" is a male singer.

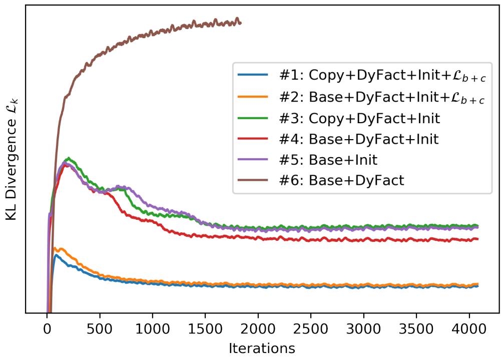


Figure 3: The Kullback–Leibler Divergence between the between the $\mathbf{z_{prior}}$ and $\mathbf{z_{post}}$ on Chinese Weibo against the training iteration number.

## 5 Conclusion and Future Work

To bridge the gap of the knowledge between machines and human beings in the dialogue generation, this paper proposes a novel knowledge-aware model ConKADI. The proposed Felicitous Fact mechanism can help the ConKADI focus on the facts that are highly relevant to the dialogue context, by generating a felicitous fact probability distribution over the retrieved facts. Besides, the proposed Context-Knowledge Fusion and Flexible Mode Fusion can facilitate the integration of the knowledge in the ConKADI. Extensive evaluations over both an open-released English dataset and our constructed Chinese dataset demonstrate our ConKADI can significantly outperform the state-of-the-art model CCM and other baselines in most experiments.

Although ConKADI has achieved a notable performance, there is still much room to improve. 1) While ATS2S$_{MMI}$ is behind our ConKADI, we find MMI can effectively enhance the ATS2S; hence, in the future, we plan to verify the feasibility of the re-ranking technique for knowledge-aware models. 2) We will continue to promote the integration of high-quality knowledge, including more types of knowledge and a more natural integration method.

## Acknowledgments

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

Our deepest gratitude goes to the reviewers for their thoughtful suggestions, and we need to thank all our team members in FineLab for their help.

## References

Antoine Bordes, Nicolas Usunier, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-Relational Data. In *Advances in NIPS*, volume 26, pages 2787–2795.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A Survey on Dialogue Systems: Recent Advances and New Frontiers. *ACM SIGKDD Explorations Newsletter*, 19.

Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *empirical methods in natural language processing*, pages 1724–1734.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2016. Fast and accurate deep network learning by exponential linear units (elus). In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, and Shuming Shi. 2019. Generating multiple diverse responses for short-text conversation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6383–6390.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5110–5117.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. 2016. Incorporating Copying Mechanism in Sequence-to-Sequence Learning.

Pei Ke, Jian Guan, Minlie Huang, and Xiaoyan. 2018. Generating Informative Responses with Controlled Sentence Function. *Proceedings of ACL*, pages 1499–1508.

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

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A Diversity-Promoting Objective Function for Neural Conversation Models. *north american chapter of the association for computational linguistics*, pages 110–119.

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

Juntao Li and Rui Yan. 2018. Overview of the NLPCC 2018 shared task: Multi-turn human-computer conversations. In *Natural Language Processing and Chinese Computing - 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26-30, 2018, Proceedings, Part II*, pages 446–451.

Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael 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.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018 , Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1489–1498.

Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2019. Knowledge aware conversation generation with explainable reasoning over augmented graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1782–1792, Hong Kong, China. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. *empirical methods in natural language processing*.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to Backward and Forward Sequences: A Content-Introducing Approach to Generative Short-Text Conversation. pages 3349–3358.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2016. A hierarchical latent variable encoder-decoder model for generating dialogues. *CoRR*, abs/1605.06069.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural Responding Machine for Short-Text Conversation. In *Annual Meeting of the Association for Computational Linguistics*, pages 1577–1586.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451.

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

Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao. 2017. How to Make Context More Useful? An Empirical Study on Context-Aware Neural Conversational Models. pages 231–236.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer Networks.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *Computer Science*.

Sixing Wu, Dawei Zhang, Ying Li, Xing Xie, and Zhonghai Wu. 2018. HL-EncDec: A Hybrid-Level Encoder-Decoder for Neural Response Generation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 845–856.

Yu Wu, Wei Wu, Dejian Yang, Can Xu, Zhoujun Li, and Ming Zhou. 2017. Neural Response Generation with Dynamic Vocabularies.

Can Xu, Wei Wu, Chongyang Tao, Huang Hu, Matt Schuerman, and Ying Wang. 2019. Neural response generation with meta-words. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019 , Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5416–5426.

Lili Yao, Yaoyuan Zhang, Yansong Feng, Dongyan Zhao, and Rui Yan. 2017. Towards implicit content-introducing for generative short-text conversation systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2190–2199.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18) , New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4970–4977.

Tiancheng Zhao, Kyusong Lee, and Maxine Eskénazi. 2018. Unsupervised discrete sentence representation learning for interpretable neural dialog generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018 , Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 654–664.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense Knowledge Aware Conversation Generation with Graph Attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 4623–4629, California. International Joint Conferences on Artificial Intelligence Organization.

Qingfu Zhu, Lei Cui, Weinan Zhang, Furu Wei, and Ting Liu. 2019. Retrieval-enhanced adversarial training for neural response generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019 , Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3763–3773.

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