Improving Knowledge-Aware Dialogue Response Generation by Using Human-Written Prototype Dialogues

Sixing Wu[1] , Ying Li[2*] , Dawei Zhang[1] 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
{wusixing, li.ying, daweizhang}@pku.edu.cn
zhwu@ss.pku.edu.cn

Abstract

Incorporating commonsense knowledge can alleviate the issue of generating generic responses in open-domain generative dialogue systems. However, selecting knowledge facts for the dialogue context is still a challenge. The widely used approach Entity Name Matching always retrieves irrelevant facts from the view of local entity words. This paper proposes a novel knowledge selection approach, Prototype-KR, and a knowledge-aware generative model, Prototype-KRG. Given a query, our approach first retrieves a set of prototype dialogues that are relevant to the query. We find knowledge facts used in prototype dialogues usually are highly relevant to the current query; thus, Prototype-KR ranks such knowledge facts based on the semantic similarity and then selects the most appropriate facts. Subsequently, Prototype-KRG can generate an informative response using the selected knowledge facts. Experiments demonstrate that our approach has achieved notable improvements on the most metrics, compared to generative baselines. Meanwhile, compared to IR(Retrieval)-based baselines, responses generated by our approach are more relevant to the context and have comparable informativeness.

1 Introduction

Unlike human beings, generative dialogue systems tend to generate generic responses, such as 'I don't know.' (Li et al., 2016). One possible reason is the gap in utilizing background knowledge. Human beings can naturally frame their dialogue understanding and responding with various learned background knowledge during the conversation. However, traditional dialogue systems can merely access the surface knowledge in the given query (Ghazvininejad et al., 2018). To tackle this issue, a feasible scheme is incorporating external knowledge into the dialogue generation (Qin et al., 2019; Wu et al., 2020b). This paper focuses on introducing the structured open-domain commonsense knowledge graph into the single-turn dialogue response generation. Commonsense knowledge refers to the widely-used everyday knowledge, for example, 'lemon tastes sour'.

In general, a knowledge graph can be regarded as a set of (e_{head}, r, e_{tail}) fact triplets. For the knowledge-aware dialogue generation, the first step is knowledge selection, aiming at selecting appropriate knowledge facts for the current dialogue context. Traditional works (Zhou et al., 2018) always adopt the Entity Name Matching (ENM), i.e., knowledge facts are retrieved based on the entity words that appear in the given query. For example, the fact triplet (apple, IsATypeOf, fruit) can be selected for the query 'What's your favourite fruit?'. Although such a widely-used method works to some extent, it has several flaws. **First**, only 1-hop knowledge can be retrieved. **Second**, instead of using the utterance-level (global) features, it uses local words to retrieve; thus, irrelevant knowledge facts may be selected. **Third**, vertex (entity) degrees in a graph are always unequal; hence, once an entity in the query corresponds to a hot vertex, the number of selected facts can be tremendous. For the time efficiency, in the practical dialogue generation, we always have an upper bound to restrict the number of involved facts. Consequently, a fact may be randomly discarded, no matter it is a highly relevant fact or an irrelevant fact; because ENM can't judge the relevance of a retrieved fact.

As shown in Table 1, to address such issues, this paper proposes a novel knowledge selection approach, Prototype-KR, which retrieves high-quality knowledge facts from prototype dialogues. Prototype dialogues are a set of diverse, informative, and knowledgeable human-written dialogues, which can be retrieved from a large-scale dialogue reposi-

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1402–1411
November 16 - 20, 2020.

Query	Oh, my **phone** is already broken.
Prototype's Query	Your old **phone** is broken?
Prototype's Response	Yes, so I bought a new **iPhone**.
Prototype's Knowledge	(**Phone, related_to, iPhone**)
Generated Response	It's time to buy a new **iPhone**.

Table 1: An example of our approach. For a query, Prototype-KR retrieves relevant prototype dialogues from the repository using an IR system, then ranks and selects the used knowledge facts from prototype dialogues. Subsequently, Prototype-KRG generates a new response based on the selected knowledge facts.

tory. Previous studies (Wu et al., 2019; Cai et al., 2019) have shown that prototype dialogues always are highly relevant to the current dialogue context; thus, Prototype-KR assumes knowledge facts that are used in the prototype dialogues would be similarly relevant to the current dialogue context. The methodology can be summarized as 1) Prototype-KR first retrieves prototype dialogues that are semantically relevant to the given query using an IR (Information Retrieval) system; 2) Prototype-KR extracts all used facts from prototype dialogues; 3) Prototype-KR selects the most appropriate knowledge facts by ranking; 4) Finally, Prototype-KRG generates a response using the knowledge facts retrieved from both the Entity Name Matching and the Prototype-KR.

Our experiments are conducted on a large-scale Chinese conversation dataset (Li and Yan, 2018) and a widely used commonsense knowledge graph ConceptNet. The experimental results demonstrate our approach outperforms both generative baselines and IR-based baselines. We also conduct a series of extensive experiments to analyze the Prototype-KR. We find our Prototype-KR can retrieve higher-quality knowledge facts compared to the traditional Entity Name Matching.

Our contributions can be summarized as 1)We propose a new knowledge selection approach, Prototype-KR, which uses prototype dialogues to effectively alleviate the flaws of the traditional approach Entity Name Matching; 2)We propose a knowledge-aware dialogue model, Prototype-KRG, for improving the knowledge-aware dialogue generation; 3) Extensive experiments empirically verify the effectiveness of our approaches.

2 Related Work

Dialogue Systems: Roughly, dialogue systems can be classified as either retrieval-based systems or generative systems (Chen et al., 2017). For generative systems, dialogue generation is always modeled as a Seq2Seq problem (Sutskever et al., 2014; Vinyals and Le, 2015). Generally, an Encoder summarizes the given query into intermediate representations, and a Decoder uses them to generate a response. Traditional methods suffer from generating generic responses, decreasing the interest of end-users. To make the dialogue more diverse and informative, previous studies have tried a lot from multiple aspects. For example, using new training objective (Li et al., 2016), using latent variables (Zhao et al., 2018; Gao et al., 2019), introducing content words (Yao et al., 2017; Xu et al., 2019).

Knowledge-Aware Methods: One crucial factor that causes generating boring responses is the insufficiency of background knowledge. Traditional models can merely access the surface knowledge from the plain text of the query (Ghazvininejad et al., 2018). Researchers have shown the generated dialogue responses can be more diverse and informative, by introducing the external knowledge, such as the unstructured background documents (Meng et al., 2019), structured knowledge graphs (Zhou et al., 2018; Wu et al., 2020a) and knowledge tables (Qin et al., 2019), or the hybrid of them (Liu et al., 2019).

Knowledge Selection: For the knowledge-aware dialogue generation, selecting appropriate knowledge facts from the knowledge graph for a specific dialogue context is still a challenge. As mentioned, the traditional Entity Name Matching has many flaws, and thus many efforts have been devoted to enhancing this knowledge selection process. (Liu et al., 2019) adopts a neural knowledge reasoning network to select an appreciate fact. (Wang et al., 2019) transfers question representation and knowledge matching abilities from KBQA systems. Although such works have achieved promising results, they always are not wise choices in the practical scenario. First, such approaches adopt complicated external networks to select knowledge, which would significantly increase parameters and make the training/inference more time-consuming. Next, the external networks require a large amount of additional labeled data, which may not be an easy thing in practice. Our work differs from them in

that: 1) Our approach does not require any additional data. Prototype dialogues can be retrieved from the training corpus. 2) Our IR-based knowledge selection is fast and requires no pre-training.

Prototypes: Recently, the research of prototype dialogues has received much attention in the context of dialogue generation owing to its high-quality. (Weston et al., 2018) encodes a retrieved prototype dialogue into vectors, and then regards them as additional features to help the dialogue generation. (Wu et al., 2019) generates a response by editing a prototype response. (Cai et al., 2019) proposes a two-step skeleton-based dialogue generation. A notable shortage of such methods is they often use only one prototype dialogue (Tian et al., 2019); thus, if the given prototype dialogue is irrelevant to the context, the generation quality will sharply decrease. Besides, such methods sometimes may degenerate to directly copy the prototype response, rather than selectively extract useful information. In contrast to these works: 1) Prototype-KR can utilize multiple prototypes at the same time; 2) Prototype-KRG is a fully generative approach, the dialogue generation process does not rely on copying or editing prototype dialogues.

3 Approach

3.1 Problem Formulation and Overview

Let $\mathcal{D} = \{(X_i, Y_i)\}^{|\mathcal{D}|}$ be a dialogue corpus, $\mathcal{K} = \{f_j\}^{|\mathcal{K}|}$ be a knowledge graph, where X is a query, Y is a response, and $f = (e_{head}, r, e_{tail})$ is a knowledge fact triplet. The prototype dialogue repository $\mathcal{D}^{'}$ can be either the $\mathcal{D}$ or a new corpus. Prototype-KR retrieves a set of prototype dialogues $\mathcal{S}^{'} = \{(X_i^{'}, Y_i^{'})\}$ from $\mathcal{D}^{'}$, and extracts all used facts from $\mathcal{S}^{'}$ (denoted as F_{raw}^{p}). Next, Prototype-KR ranks facts $\in F_{raw}^{p}$, and selects top-k facts (denoted as F^{p}). Meanwhile, Entity Name Matching is also used to retrieve a set of facts (denoted as F^{n}) . Finally, Prototype-KRG uses F^{p}, F^{n}, and X to generate the target response: $p(Y|X, F^{p}, F^{n})$.

3.2 Prototype-KR

Prototype-KR is a 3-stage method to retrieve top-k relevant knowledge facts from prototype dialogues.

Prototype Retrieval: Prototype dialogues $\mathcal{S}^{'}$ are firstly retrieved from the repository $\mathcal{D}^{'}$. We adopt Lucene[1] to construct an index and use its built-in engine to retrieve $5k$ prototype dialogues. Following (Wu et al., 2019), we have different strategies in the training and inference. In the training, we retrieve prototype dialogues based on the response similarity; in the test, we retrieve prototype dialogues based the query similarity. For each prototype dialogue pair $(X_i^{'}, Y'_i) \in \mathcal{S}^{'}, i \in [1:5k]$, we extract all its knowledge facts to the subset $F_i^{'}$. Afterwards, all subsets are merged together: $F_{ALL}^{'} = F_1^{'} \cup \cdots \cup F_{5k}^{'}$

Coarse-Grained Ranking: For each knowledge fact $f_j \in F_{ALL}^{'}$, the corresponding coarse-grained ranking score s_j^c is computed as:

$$s_j^c = \sum_{i=1:5k} I_{i,j} \times J(P_{X/Y}, P_{X_i^{'}/Y_i^{'}}) \times IDF(f_j) \quad (1)$$

where $P_{X/Y}$/$P_{X_i^{'}/Y_i^{'}}$ refers to X/$X_i^{'}$ in the test, and refers to Y/$Y_i^{'}$ in the training, the indicator $I_{i,j}$ is 1 if $f_j \in F_i^{'}$ else 0, $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$ measures the Jaccard similarity between A and B from a bag-of-word view, and the inverse document frequency $IDF(\cdot)$ is used to penalize the high-frequency generic knowledge facts. We keep top-$2k$ ranked knowledge facts (denoted as $F_{CGR}^{'}$). It is worth noting that, for each unique target entity e_{tail}, we only keep one fact with the highest score.

Fine-Grained Ranking: For each $f_j \in F_{CGR}^{'}$, we use embedding-based metric to measure the semantic relevance to the current query/response $P_{X/Y}$ (denoted as P in the below), and then we remain top-k ranked facts (i.e., F^{p}). The corresponding fine-grained score s_j^f is computed as:

$$s_j^c \times \theta(\mathbf{E_x}(P), \mathbf{E_w}(e_{head})) \times \theta(\mathbf{E_x}(P), \mathbf{E_w}(e_{tail})) \quad (2)$$

where θ is the cosine similarity, $\mathbf{E_x}$ is the sentence-level extrema embedding. For each dimension of the word embedding vectors, take the most extreme value among all vectors in the sentence (Liu et al., 2016). $\mathbf{E_w}(e_{head/tail})$ is the pre-trained word embedding of the head/tail tail entity of f_j.

3.3 Prototype-KRG

3.3.1 Context Encoder

Context Encoder is a bi-directional GRU network (Cho et al., 2014), aiming at encoding the query X into intermediate representations. The forward GRU reads X from the beginning to the end; the backward GRU reads X from the end to the beginning. At the time step t, two outputs are given by:

[1] https://lucenenet.apache.org

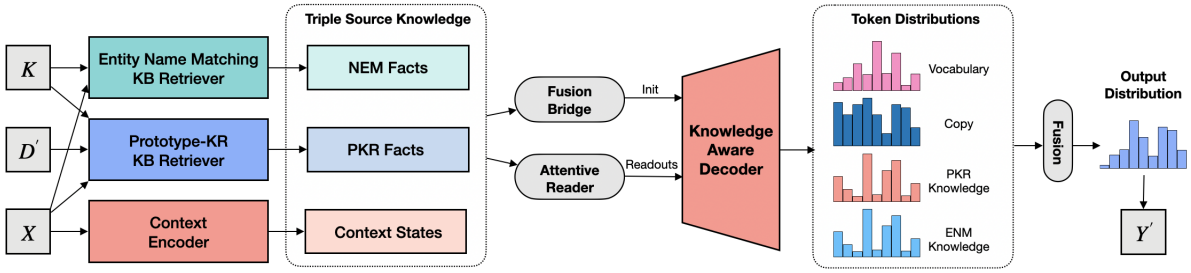

Figure 1: The architecture of our approach. $\mathcal{K}$ denotes a knowledge graph, $\mathcal{D}^{'}$ is a dialogue repository used to retrieve prototype dialogues, X is a query, and $Y^{'}$ is a generated response.

$$\mathbf{h_t^f} = GRU(\mathbf{E_v}(x_t), \mathbf{E_k}(x_t), \mathbf{h_{t-1}^f});$$

$$\mathbf{h_t^b} = GRU(\mathbf{E_v}(x_{n-t+1}), \mathbf{E_k}(x_{n-t+1}), \mathbf{h_{t-1}^b}) \quad (3)$$

where $\mathbf{E_v}(x)$ is a learn-able word embedding of x, and the corresponding fixed entity embedding $\mathbf{E_k}(x)$ is employed to augment the semantic understanding if x is an entity word. The concatenation $\mathbf{h_t} = [\mathbf{h_t^f}; \mathbf{h_t^b}]$ is regarded as the output at the time step t. Consequently, the encoded intermediate representation of X is $\mathbf{H} = (\mathbf{h_1}, \mathbf{h_2}, ..., \mathbf{h_n})$.

3.3.2 Knowledge Bridge Fusion

Context Encoder only summarizes the surface text of X. For accessing the knowledge before the generation, we propose the Knowledge Bridge Fusion, which uses both the intermediate representation $\mathbf{H}$ and the knowledge facts to initialize the Decoder. Given the last context state $\mathbf{h_n}$, We first obtain the attention $\mathbf{a^p}$ of $\mathbf{F^p}$, and the attention $\mathbf{a^n}$ of $\mathbf{F^n}$:

$$\mathbf{a^p} = KA(\mathbf{F^p}, \mathbf{h_n}) \quad \mathbf{a^n} = KA(\mathbf{F^n}, \mathbf{h_n})$$

$$KA(\mathbf{F}, \mathbf{h_n}) = \textstyle\sum_{i=1:|F|} e_i^{KA} \mathbf{W_v f_i} \quad (4)$$

$$e_i^{KA} = \frac{\exp(\mathbf{W_k f_i} \cdot \mathbf{W_q h_n})}{\exp(\sum_{j=1:|F|} \mathbf{W_k f_j} \cdot \mathbf{W_q h_n})}$$

where $\mathbf{F^{p/n}}$ is the corresponding embedding of $F^{p/n}$, and KA is an Attention function. Learn-able parameters $\mathbf{W_k}, \mathbf{W_q}, \mathbf{W_o}$ are not shared between $KA(\mathbf{F^p}, \mathbf{h_n})$ and $KA(\mathbf{F^n}, \mathbf{h_n})$.

Subsequently, $\mathbf{a^p}$, $\mathbf{a^n}$ and $\mathbf{h^n}$ are fused by a MLP, and the result is regarded as the initial state of the Decoder:

$$\mathbf{z_0} = \gamma_c \mathbf{h_n} + \gamma_p \mathbf{a^p} + \gamma_n \mathbf{a^n}$$

$$\gamma_{c,n,p} = softmax(\mathbf{W_{bridge}}[\mathbf{a^n}; \mathbf{a^p}; \mathbf{h^n}; \mathbf{ita}]) \quad (5)$$

where $[\cdot;\cdot]$ is the concatenation operation; the vector $\mathbf{ita}$ is the concatenation of interactions between the $\mathbf{a^{p/n}}$ and $\mathbf{h^n}$, which includes $\mathbf{h^n} + \mathbf{a^{p/n}}$, $\mathbf{h^n} - \mathbf{a^{p/n}}$, $abs(\mathbf{h^n} - \mathbf{a^{p/n}})$.

3.3.3 Response Generation

Decoder is another GRU network, at each decoding time step t, the hidden state $\mathbf{z_t}$ is updated as:

$$\mathbf{z_t} = GRU(\mathbf{z_{t-1}}, \mathbf{c_t}, \mathbf{c_t^p}, \mathbf{c_t^n}, \mathbf{E_k}(y_{t-1}); \mathbf{E_v}(y_{t-1})) \quad (6)$$

where y_{t-1} is the last predicted token, $\mathbf{c_t}$ is the attention of $\mathbf{H}$ (see (Luong et al., 2015) for the detail)), and $\mathbf{c_t^{p/n}}$ is the attention of $\mathbf{F^{p/n}}$ (see (Bahdanau et al., 2014) for the detail).

The tokens to be generated can be one of the following four types: words from the fixed vocabulary V, words copied from X, entity words from F^p, and entity words from F^n.

Vocabulary Words: The probability distribution $p_{v,t}$ over the fixed vocabulary V is given by:

$$p_{v,t} = softmax(\mathbf{W_{vocab}}[\mathbf{z_t}; \mathbf{u_t}]) \quad (7)$$

$$\mathbf{u_t} = [\mathbf{c_t}; \mathbf{c_t^n}; \mathbf{c_t^p}; \mathbf{E_v}(y_{t-1}); \mathbf{E_k}(y_{t-1})]$$

Copied Words: Decoder can copy a word from X, the corresponding probability distribution over the query message X is given by:

$$p_{c,t} = softmax(\mathbf{H W_{cp} z_t}) \quad (8)$$

Entity Words: Decoder can select the best-matched knowledge fact from F^p and F^n, and then copy the corresponding entity word. For F^p and F^n, we apply the same method but with different parameters to compute the probability distribution:

$$p_{p/n,t} = softmax(S(\mathbf{F^{p/n}}, \mathbf{z_t})) \quad (9)$$

$$S(\mathbf{f}_{j=1:|F|}, \mathbf{z_t}) = \mathbf{v_f^\top} \tanh(\mathbf{W_e z_t} + \mathbf{U_e f_j})$$

Mode Fusion: Following (Wu et al., 2020a), such four distributions are fused using multiple mode gates:

$$p_t = \pi_{v,t}p_{v,t} + \pi_{c,t}p_{c,t} + \pi_{p,t}p_{p,t} + \pi_{n,t}p_{n,t}$$

$$(\pi_{v,t}, \pi_{c,t}, \pi_{p,t}, \pi_{n,t}) = softmax(\mathbf{W_{mode}}[\mathbf{z_t}; \mathbf{u_t}]) \quad (10)$$

3.4 Training

Finally, the training objective is given by:

$$\mathcal{L} = \mathcal{L}_{Gen} + \mathcal{L}_{BOW} + \mathcal{L}_{Mode} \quad (11)$$

where $\mathcal{L}_{Gen} = -\sum_t \log p_t(y_t|y_{t-1:1}, X, F^p, F^n)$ is the negative log-likelihood. $\mathcal{L}_{BOW}$ is the bag-of-words loss to ensure the fluency (Zhao et al., 2017), our BOW prediction takes $\mathbf{z_0}$ as the input. $\mathcal{L}_{Mode}$ is the teach-force loss, i.e., the Cross-Entropy between the $\pi_{v/c/p/n}$ and the ground-truth 0-1 mode indicator, to help Prototype-KRG more accurate when selecting a target word from four types of words (Zhou et al., 2018) .

4 Experiments

4.1 Settings

Dataset: We adopt a large-scale Chinese conversational dataset (Li and Yan, 2018), which collected conversations from the largest Chinese SNS (weibo.com). Commonsense knowledge fact triples are collected from the ConceptNet (conceptnet.io) (Speer et al., 2017), which includes 27K entities, 26 relations, and 661K triples. Utterances that are too long (>30 words), too short (<4 words), or do not contain entity information are discarded, the remaining dialogues are randomly divided into three sets: training (847K), validation (30K), and test (30K).

Models: We select generative baselines and IR-based baselines. **S2S**: The RNN-based Seq2Seq with Attention (Sutskever et al., 2014; Luong et al., 2015); **Copy**: Copy mechanism additionally allows Seq2Seq to copy a word from the query directly (See et al., 2017); **Transformer** : Rather than RNNs, both the Encoder and the Decoder are two 6-layer Transformers (Vaswani et al., 2017), respectively; **GenDS**: A strong knowledge-aware dialogue generation baseline (Zhu et al., 2017); **CCM**: One SOTA commonsense knowledge-aware dialogue generation model, which proposes a static and a dynamic graph attention mechanism (Zhou et al., 2018); **ProtoEdit**: One SOTA IR-augmented dialogue generation model by editing the prototype response (Wu et al., 2019); **IR**: We use a predefined index to retrieve a response from the dialogue repository; meanwhile, **IR-Rerank** further adds a Jaccard-based rerank process. Especially, our Prototype-KRG (denoted as **Ours**) and ProtoEdit have variants. As mentioned, in the training, the original **Ours**$_R$ and **ProtoEdit**$_R$ retrieve prototype dialogues based on the response similarity. Differently, the variants **Ours**$_Q$ and **ProtoEdit**$_Q$ retrieve prototype dialogues based on the query similarity in the training.

Implementations: For S2S, Copy, Transformer, and our approach, we implement them based on a PyTorch Seq2Seq framework, OpenNMT (Klein et al., 2017). For GenDS, we use a Tensorflow implementation. For CCM and ProtoEdit, we use their official codes. In experiments, the vocabulary is set to 30,000, the word embedding dimension is 300, the entity/relation embedding is initialized from a 100-dimensional pre-trained embedding learned by TransE (Bordes et al., 2013), RNNs are implemented as 1024-dimensional GRUs, Adam is used to optimizing the model with an initial learning rate 0.0001 and the batch size 64; learning rate will be halved if the perplexity on the validation set starts to increase, the training will be stopped if the perplexity on the validation set increases in two successive epochs. In the inference, the beam width is set to 10. For a fair comparison, such settings are similarly applied to other implementations as possible. Under such settings, our approach has 193M parameters (including embeddings), and the training takes about 1.5 days on an Nvidia 2080Ti.

Metrics: We have multiple criteria. The first metric **EntN** measures knowledge utilization, i.e., the number of generated knowledgeable entities per generated response (Zhou et al., 2018). For the relevance, we employ two embedding based metrics, Embedding-Greedy (**EmG**), Embedding-Extrema (**EmX**), and two word-overlap-based metrics, **ROUGE**, **BLEU-4** (Liu et al., 2016). Next, we measure the diversity by reporting the ratio of distinct uni/bi-grams (**DIST1/2**) in all generated words (Li et al., 2016). Lastly, **Entropy** is used to measure the informativeness (Mou et al., 2016). Meanwhile, to illustrate the overall performance, we design two auxiliary metrics **Overall**$_{+DI}$ and **Overall**. For a model, we take S2S as the stan-

Type	Method	EntN	EmG	EmX	ROUGE	BLEU4	DIST1	DIST2	Entropy	Overall$_{+DI}$	Overall
Gen	S2S	0.92	0.590	0.633	9.16	0.36	0.90	4.29	7.47	1.00	1.00
Gen	Copy	0.98	0.599	0.636	9.15	0.36	2.85	9.96	7.76	1.45	1.02
Gen	Transformer	0.61	0.558	0.613	6.47	0.19	2.78	8.81	**8.75**	1.27	0.83
Gen	GenDS	1.01	0.603	0.643	11.53	0.44	2.22	10.37	7.65	1.44	1.10
Gen	CCM	1.13	0.612	0.644	12.67	0.78	1.11	5.33	6.60	1.27	1.28
Gen	**Ours**$_Q$	1.20	**0.629**	**0.655**	**12.96**	**0.90**	2.67	13.56	8.05	1.81	1.39
Gen	**Ours**$_R$	**1.43**	0.627	0.650	12.42	0.84	**3.60**	**20.32**	8.30	**2.15**	**1.40**
IR	ProtoEdit$_Q$	0.64	0.612	0.638	9.98	0.54	0.44	2.77	7.12	0.93	1.04
IR	ProtoEdit$_R$	0.68	0.587	0.623	7.42	0.27	2.00	23.52	8.17	1.63	0.89
IR	IR	0.71	0.574	0.624	6.31	0.26	8.23	49.20	9.99	3.27	0.91
IR	IR-Rerank	0.85	0.574	0.623	6.37	0.29	7.90	47.60	9.86	3.20	0.95

Table 2: Automatic evaluation results. Considering the difference between IR-based and generative systems, we compare different types of the model separately: scores in **bold** stand for the leadership among generative models; scores with an underline stand for the leadership among our models and IR-based models.

dard; then, we calculate out the relative scores to the S2S metric by metric, and the averaged relative score is **Overall**$_{+DI}$. IR-based methods can access human-written dialogues, which brings them additional advantages in diversity and informativeness. It would be better to exclude such metrics into the overall score when comparing across generative methods and IR-based methods. Hence, the calculation of **Overall** excludes such metrics.

4.2 Experimental Results

Experimental results have been reported in Table 2.

vs. Generative Baselines: Prototype-KRG outperforms generative baselines in terms of most metrics and the overall scores. Prototype-KRG only slightly loses the leadership in terms of Entropy compared to the Transformer. The advantages of the previous SOTA CCM and our Prototype-KRG show that knowledge can indeed help the dialogue generation. Compared with two knowledge-aware baselines, GenDS and GenDS, Prototype-KRG is notably better in terms of the knowledge utilization, diversity, and the informativeness. It can be attributed to 1) Prototype-KR can select higher-quality knowledge facts; 2) the effectiveness of Prototype-KRG.

vs. IR-based Baselines: Generative approaches are not directly compared with IR-based approaches, because of their different characteristics. The later type naturally has higher diversity and informativeness, for they can directly access the human-written dialogues. Thus, IR and IR-Rerank significantly outperform other models in terms of the DIST-1/2 and the Entropy. However, every coin has two sides; they suffer from low relevance; they are notably beaten by generative approaches in the relevance metrics. This is because they mechanically return existing unmodified dialogues even when the retrieved responses are irrelevant to the query. ProtoEdit tries to address this flaw by editing the retrieved dialogue. It can be seen that diversity and informativeness have significantly decreased, but the improvement of the relevance and the overall performance (see **Overall**) is still limited. Compared to ProtoEdit, Prototype-KRG has comparable performance in terms of diversity, and notably better performance in the remaining aspects and the overall performance.

How to Select Prototypes: As mentioned, there are two strategies to retrieve prototype dialogues in training. We have noticed that the authors of ProtoEdit said that ProtoEdit$_Q$ always generates non-sense responses (Wu et al., 2019). As reported in Table 2 and our manually reviewing, compared to ProtoEdit$_R$, responses generated by ProtoEdit$_Q$ are indeed boring and non-sense, while more relevant to the query. Unlike ProtoEdit, although Ours$_R$ similarly outperforms Ours$_Q$ in terms of DIST1/2 and Entropy, such two implementations are comparable in the aspect of the relevance (see **Overall**). It means our approach is much more robust to different prototype dialogues.

Human Annotation: We employed three annotators and sampled 200 queries from the test. Six baselines (1200 pairs) are involved in our pairwise comparisons. There are two criteria: (1) Appropriateness (i.e., fluency and relevance); (2) Informativeness (how much relevant knowledge is provided). The agreement among annotators are: for the appropriateness, 2/3 agreement is 97.3%, 3/3 agreement is 54.7%; for the informativeness, 2/3 agreement is 97.6%, 3/3 agreement is 55.0%. As shown in Table 3, our approach outperforms all baselines, indicating the advantage of our ap-

proaches. In terms of appropriateness, S2S and Copy are the two best baselines because they always generate generic responses, which are fluent and sometimes are easy to be accepted by humans. CCM performs poorly because it sometimes generates long but not fluent responses. Two IR-based methods are unsatisfactory. Responses given by ProtoEdit and IR-Rank are fluent, but sometimes irrelevant to the query. Moving to the informativeness, CCM is the best generative baseline, which indicates the importance of using knowledge. Benefit from accessing the human-written dialogues, IR-based ProtoEdit and IR-Rank outperform generative baselines. If we ignore the dialogue context and only check the informativeness of responses, IR-Rerank can outperform ProtoEdit and ProtoEdit has comparable performance with our approach. However, the context should be considered, and thus we penalized the irrelevant information; as a result, ProtoEdit is comparable with IR-Rerank, and our approach is better than ProtoEdit.

vs.	$\mathbf{A}_+$	$\mathbf{A}_0$	A_-	$\mathbf{I}_+$	$\mathbf{I}_0$	$\mathbf{I}_-$
S2S	**.515**	.178	.307	**.630**	.140	.230
Copy	**.527**	.155	.318	**.650**	.112	.238
GenDS	**.580**	.152	.268	**.645**	.117	.238
CCM	**.667**	.088	.245	**.580**	.072	.348
ProtoEdit$_R$	**.595**	.098	.307	**.512**	.080	.408
IR-Rerank	**.622**	.167	.211	**.500**	.142	.358

Table 3: Human annotation results. **A/I** is **Appropriateness/Informativeness**. $+/0/-$ means Ours$_R$ wins/ties/loses the comparison. Our approach is better than baselines (sign test, p-value < 0.005).

4.3 Analysis of Prototype-KR

Ablation Study: Following (Zhou et al., 2018; Wang et al., 2019) and many other works, in the above experiments, at least one golden fact used by the reference response[2] is given in the test set. To clearly illustrate the difference among variants, we construct a new test set in line with the practice. Instead of manually assuring such a golden knowledge fact is existing, we do not add any additional fact. As shown in Table 4, compared to the 'Full', although the variant '-Dual' similarly uses the knowledge facts retrieved by PKR and ENM, and it has similar performance in the aspect of the relevance, we find the metric EntN and DIST2 have significantly decreased, indicating the necessity to distinguish them in a model. Next, we turn to compare PKR and ENM; we find '-PKR' impacts the performance more than '-ENM', which illustrates the knowledge quality of PKR is better than ENM's. '-PKR-ENM' removes the use of knowledge, and then, most metrics drop dramatically, which indicates the importance of introducing knowledge. In summary 1) Knowledge is indispensable in the dialogue generation; 2) Our Prototype-KR can select more appropriate knowledge facts than the traditional Entity Name Matching.

[2]Reference responses are used to evaluate generated responses.

Config	EntN	EmG	BLEU4	DIST2	Overall
Full	1.42	0.62	0.72	20.35	1.31
-Dual	1.30	0.62	0.74	14.64	1.29
-PKR	1.00	0.61	0.74	15.34	1.23
-ENM	1.49	0.61	0.65	21.63	1.28
-PKR-ENM	0.98	0.60	0.36	9.96	1.02

Table 4: Ablation tests. 'Full' is **Ours**$_R$, '-PKR/-ENM' excludes knowledge retrieved by the Prototype-KR/Entity Name Matching, '-Dual' does not distinguish facts from PKR/NEM, facts are mixed together.

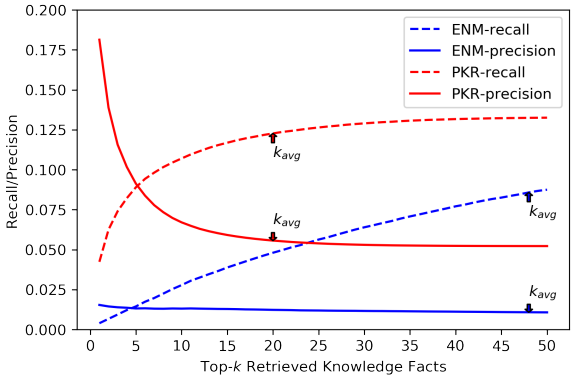

Figure 2: For the top-k retrieved knowledge facts, we calculate the recall $\frac{|E_Y \cap E_F|}{|E_Y|}$ and precision $\frac{|E_Y \cap E_F|}{|E_F|}$, where E_Y are entities that are used in the reference response, and E_F are entities in the retrieved facts; $k_{avg} = 19.5/48.2$ masks the average length of E_F.

Knowledge Selection: Figure 2 reports the statistical evaluations for the knowledge facts retrieved by PKR and NEM. We have several observations: 1) Both two metrics indicate that PKR notably outperforms NEM in knowledge selection; 2) PKR has more notably advantages when k is small, which means the ranking of PKR is accurate, highly relevant facts always have prior ranks.

Case Study: Table 5 reports three examples. In the **first example**, although all approaches have generated fluent responses, they are different in both appropriateness and informativeness. S2S and Copy generated generic responses. For the

#1 Query: 哦哦，自恋无罪，世界很美。 Oh, narcissism is innocent, the world is beautiful.
S2S: 我也喜欢。 I love it,too. Copy: 我喜欢。 I love it. GenDS: 不不不,你是自信。No, no, no, you're confident. CCM: 我是自信。 I am confident. ProtoEdit: 喜欢就好。 I'm glad you love it. IR: 莉莉喜欢就好。 I'm glad Lily loves it. Ours: 生命在于自信。 Life is about confidence.
#2 Query: 蓝色那个什么花? What is that blue flower?
S2S: 蓝色。 Blue. Copy: 蓝色的。 Blue flowers. GenDS: 叶子。Leaves. CCM: UNK 的叶子。 The leaves of UNK. ProtoEdit: 我是最喜欢的花。 I'm my favorite flower. IR: 哼那个蓝丑丑是我最喜欢的。 Hum, that blue clown is my favorite. Ours: 路边的野花。Wild flowers on the road.
#3 Query: 好漂亮，这些都是什么品种的花? How beautiful, what kind of flowers are these?
rank-1: (漂亮 Beautiful, Synonym, 好看 Good-Looking) rank-2: (郁金香 Tulip, IsA, 花 Flower) rank-3: (花 Flower, SymbolOf, 漂亮 Beautiful) Ours: 都是郁金香。There are tulips

Table 5: Case study.

knowledge-aware GenDS and CCM, they detected a specific topic (self-confidence), but the generated responses are a little irrelevant to the query. Similarly, ProtoEdit and IR answered two generic responses. In the **second example**, S2S and Copy repeated the words, GenDS and CCM used a wrong knowledge ('flower' is not 'leaf'). ProtoEdit and IR gave two weird responses. The **last example** first shows top-3 knowledge facts that were retrieved and ranked by our Prototype-KR, and then shows the response generated by Prototype-KRG. It can be seen that such three facts are highly relevant to the query, and the generated response uses the second fact. In short, compared to the generative approaches, our approach can generate more informative and relevant responses; compared to the IR-based approaches, our approach can generate more relevant responses.

Error Analysis: We have further labeled the error type for 200 responses sampled in the above human annotation. For a response, it can be labeled as a **perfect** (beyond the expectation), a **good** (acceptable), or a **bad** response. For a bad case, we give it **one or more** fine-grained error types. There are five error types: being **irrelevant** to the dialogue context, including **illusory** errors or **grammar** errors, generating some **repeated** words, and **non-sense**. About 64.5% generated responses are labeled as perfect or good; the remaining 35.5% more or less have some mistakes. The most notable error type is 'non-sense', which means the generated response is meaningless while it is always fluent and relevant to the context; for example, wrongly rephrases the query. Responding with an irrelevant topic, making grammar errors, and repeating words are three common error types among generative models, but their error rates in our approach are well-controlled. Knowledge-aware models are more potential to generate 'illusory' responses that violate the commonsense knowledge, for example, 'What disease do you drink?'. We are glad to find this rarely happens in our approach.

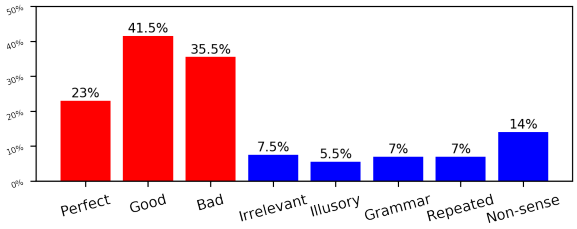

Figure 3: The statistics of error type. Red bars are exclusive labels; each response can only be labeled as one type. Blue bars are fine-grained error type labels; each bad case is given at least one label.

5 Conclusion and Future Work

We propose a novel knowledge selection method, Prototype-KR, and a knowledge-aware model, Prototype-KRG, for the open-domain knowledge-aware dialogue generation. Prototype-KR retrieves knowledge facts from the human-written prototype dialogues, which is fast, accurate and requires no additional labeled data. Extensive experiments on a large-scale Chinese dataset show that our approach outperforms generative baselines on most metrics. Compared to IR-based approaches, our approach can generate responses with higher relevance and comparable informativeness.

In the future, we will continue to strengthen the use of prototype dialogues without making the dialogue generation process complicated. Meanwhile, we are going to research the possibility to use different knowledge in a dialogue system.

Acknowledgments

This work is supported by the National Key R&D Program of China (Grant No. 2017YFB1002000).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation By Jointly Learning To Align and Translate. In *International Conference on Learning Representations*, pages 1–15.

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

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi. 2019. Skeleton-to-response: Dialogue generation guided by retrieval memory. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1219–1228. Association for Computational Linguistics.

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, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, pages 1724–1734.

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

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 *AAAI*, pages 5110–5117.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

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

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

Angli Liu, Jingfei Du, and Veselin Stoyanov. 2019. Knowledge-augmented language model and its application to unsupervised named-entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1142–1150, Minneapolis, Minnesota. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian 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. In *EMNLP*, pages 2122–2132.

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

Chuan Meng, Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2019. Refnet: A reference-aware network for background based conversation. *CoRR*, abs/1908.06449.

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. In *COLING*, pages 3349–3358.

Libo Qin, Yijia Liu, Wanxiang Che, Haoyang Wen, Yangming Li, and Ting Liu. 2019. Entity-consistent end-to-end task-oriented dialogue system with KB retriever. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 133–142. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. 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 1073–1083. Association for Computational Linguistics.

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

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

Zhiliang Tian, Wei Bi, Xiaopeng Li, and Nevin L. Zhang. 2019. Learning to abstract for memory-augmented conversational 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 3816–3825. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

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

Jian Wang, Junhao Liu, Wei Bi, Xiaojiang Liu, Kejing He, Ruifeng Xu, and Min Yang. 2019. Improving knowledge-aware dialogue generation via knowledge base question answering. *CoRR*, abs/1912.07491.

Jason Weston, Emily Dinan, and Alexander H. Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. In *Proceedings of the 2nd International Workshop on Search-Oriented Conversational AI, SCAI@EMNLP 2018, Brussels, Belgium, October 31, 2018*, pages 87–92. Association for Computational Linguistics.

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2020a. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5811–5820, Online. Association for Computational Linguistics.

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2020b. Topicka: Generating commonsense knowledge-aware dialogue responses towards the recommended topic fact. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3766–3772.

Yu Wu, Furu Wei, Shaohan Huang, Yunli Wang, Zhoujun Li, and Ming Zhou. 2019. Response generation by context-aware prototype editing. 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 7281–7288.

Can Xu, Wei Wu, Chongyang Tao, Huang Hu, Matt Schuerman, and Ying Wang. 2019. Neural response generation with meta-words. In *ACL*, 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 *EMNLP*, pages 2190–2199.

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

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL*, 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 *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. *CoRR*, abs/1709.04264.