KdConv: A Chinese Multi-domain Dialogue Dataset Towards Multi-turn Knowledge-driven Conversation

Hao Zhou*, Chujie Zheng*, Kaili Huang, Minlie Huang†, Xiaoyan Zhu
Department of Computer Science and Technology,
Institute for Artificial Intelligence, State Key Lab of Intelligent Technology and Systems,
Beijing National Research Center for Information Science and Technology,
Tsinghua University, Beijing 100084, China
{tuxchow,chujiezhengchn}@gmail.com, aihuang@tsinghua.edu.cn

Abstract

The research of knowledge-driven conversational systems is largely limited due to the lack of dialog data which consists of multi-turn conversations on multiple topics and with knowledge annotations. In this paper, we propose a Chinese multi-domain knowledge-driven conversation dataset, **KdConv**, which grounds the topics in multi-turn conversations to knowledge graphs. Our corpus contains 4.5K conversations from three domains (film, music, and travel), and 86K utterances with an average turn number of 19.0. These conversations contain in-depth discussions on related topics and natural transition between multiple topics. To facilitate the following research on this corpus, we provide several benchmark models. Comparative results show that the models can be enhanced by introducing background knowledge, yet there is still a large space for leveraging knowledge to model multi-turn conversations for further research. Results also show that there are obvious performance differences between different domains, indicating that it is worth further explore transfer learning and domain adaptation. The corpus and benchmark models are publicly available[1].

1 Introduction

It has been a long-term goal of artificial intelligence to deliver human-like conversations, where background knowledge plays a crucial role in the success of conversational systems (Shang et al., 2015; Li et al., 2016a; Shao et al., 2017). In task-oriented dialog systems, background knowledge is defined as slot-value pairs, which provides key information for question answering or recommendation, and has been well defined and thoroughly studied (Wen et al., 2015; Zhou et al., 2016). In open-domain conversational systems, it is important but challenging to leverage background knowledge, which is represented as either knowledge graphs (Zhu et al., 2017; Zhou et al., 2018a) or unstructured texts (Ghazvininejad et al., 2018), for making effective interactions.

Recently, a variety of knowledge-grounded conversation corpora have been proposed (Zhou et al., 2018b; Dinan et al., 2018; Moghe et al., 2018; Moon et al., 2019; Wu et al., 2019; Liu et al., 2018; Tuan et al., 2019; Qin et al., 2019) to fill the gap where previous datasets do not provide knowledge grounding of the conversations (Godfrey et al., 1992; Shang et al., 2015; Lowe et al., 2015). CMU DoG (Zhou et al., 2018b), India DoG (Moghe et al., 2018), and Wizard of Wikipedia (Dinan et al., 2018) demonstrate attempts for generating informative responses with topic-related Wikipedia articles. However, these datasets are not suitable for modeling topic transition or knowledge planning through multi-turn dialogs based on the relations of topics. OpenDialKG (Moon et al., 2019) and DuConv (Wu et al., 2019) use knowledge graphs as knowledge resources. Nevertheless, the number of topics is limited to one (Moon et al., 2019) or two (Wu et al., 2019), which is not sufficient for diversified topic transition in human-like conversations. Therefore, these knowledge-grounded dialog datasets still have limitations in modeling knowledge interactions[2] in multi-turn conversations.

In this paper, we propose **KdConv**, a Chinese multi-domain dataset towards multi-turn **K**nowledge-**d**riven **Conv**ersation, which is suitable for modeling knowledge interactions in multi-turn human-like dialogues, including knowledge planning, knowledge grounding, knowledge adaptations, etc. KdConv contains 86K utterances and

* Equal contribution
† Corresponding author: Minlie Huang.
[1] https://github.com/thu-coai/KdConv

[2] Refer to knowledge planning, knowledge grounding, knowledge adaptations in dialog systems.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7098–7108
July 5 - 10, 2020.

Conversation (Music)	Knowledge Triple		
	Head Entity	Relation	Tail Entity
User1: 知道《***飞得更高***》这首歌吗？ Do you know the song ‘***Flying Higher***’?			
User2: 知道呀，这首歌入选了中歌榜中国年度最受华人欢迎十大金曲。 Yes, this song has been selected in the top ten most popular songs in China.	***Flying Higher***	Information	… selected in the top ten most popular songs in China…
…			
User1: 具体的发行时间你记得吗？ Do you remember the exact release date?			
User2: 记得，是在 2005 年 3 月 19 日。 Yes. It is March 19, 2005.	Flying Higher	Release date	March 19, 2005
User1: 我觉得这首歌算是***汪峰***的经典之曲。 I think it is one of the classic songs of ***Wang Feng***.	Flying Higher	Original singer	***Wang Feng***
User2:我也那么认为，编曲填词都由他自己完成，真的算是经典之作了。 So do I. The arrangement and lyrics of the music are all completed by himself. It’s really a classic.	Flying Higher	Arrangment Lyrics	***Wang Feng***
User1: 说到他真的很了不起，在音乐方面获得很多大奖，我能说上来的就有第 12 届音乐风云榜年度最佳男歌手奖。 He is really amazing and has won many awards in music, such as the 12th Music Awards of the Year Award for Best Male Singer.	***Wang Feng***	Main achievements	The 12th Music Awards of the Year Award for Best Male Singer
…			
User1: 那他的歌曲除飞得更高，你还喜欢哪首？ So which song do you like besides ‘Flying Higher’?			
User2: 再喜欢的就是《***怒放的生命***》这首歌了，听的感觉特别好，减压。 I like ‘***Blooming Life***’. I feel great and decompression.	Wang Feng	Representative works	Fireworks, Brave Heart, Flying Higher, ***Blooming Life***…
User1: 啊，这首歌我也很喜欢，也都是由他自己作词作曲并演唱。 Oh, I like this song, too. He wrote and sang it by himself.	***Blooming Life***	Information	‘Blooming Life’ is a song sung, written and composed by Wang Feng… The song won the Best Song of the Year Award in the 13th Global Chinese Music List.
User2: 是的，该曲也获得了 13 届全球华语音乐榜中榜年度最佳歌曲奖。 Yes, and the song also won the Best Song of the Year Award in the 13th Global Chinese Music List.	***Blooming Life***	Information	‘Blooming Life’ is a song sung, written and composed by Wang Feng… The song won the Best Song of the Year Award in the 13th Global Chinese Music List.

Knowledge Graph

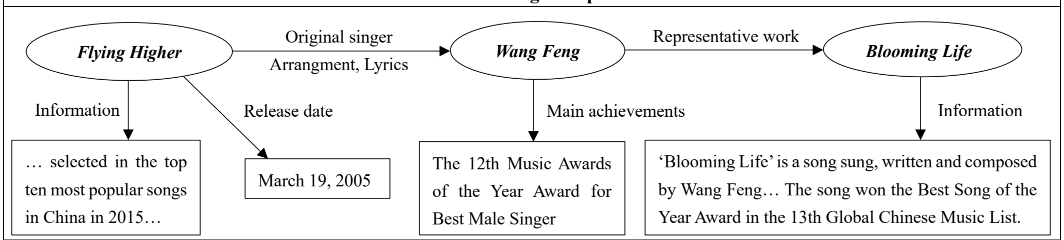

Figure 1: An example in KdConv from the **music** domain. The underlined text is the related knowledge that is utilized in conversation. The ***italic*** text and circles are topics (refer to the distinct head entities in the knowledge triples and the central nodes with degree greater than 1 in the knowledge graph) in this dialogue.

4.5K dialogues in three domains, 1.5K dialogues for each domain (an example is shown in Figure 1). Each utterance is annotated with related knowledge facts in the knowledge graph, which can be used as supervision for knowledge interaction modeling. Furthermore, conversations of KdConv contain diversified topics ranged from one to four, without any pre-defined goals or constraints, which are closer to real human-human conversations than other datasets. The relations of topics are explicitly defined in the knowledge graph. Moreover, KdConv covers three domains, including film, music, and travel, which can be used to explore knowledge adaptation between different domains. We provide a benchmark to evaluate both generation- and retrieval-based conversational models on the proposed dataset with/without access to the corresponding knowledge. Results show that knowledge grounding contributes to the improvement of these models, while existing models are still not strong enough to deliver knowledge-coherent conversations, indicating a large space for future work.

In summary, this paper makes the following contributions:

- We collect a new dataset, KdConv, for knowledge-driven conversation generation in Chinese. KdConv contains 86K utterances and 4.5K dialogues in three domains (film, music, and travel). The average turn number is about 19, remarkably longer than those in other corpora.

Dataset	Language	Knowledge Type	Annotation Level	Domain	Avg. # turns	Avg. # topics	# uttrs
CMU DoG	English	Text	Sentence	Film	22.6	1.0	130K
WoW	English	Text	Sentence	Multiple	9.0	2.0	202K
India DoG	English	Text & Table	Sentence	Film	10.0	1.0	91K
OpenDialKG	English	Graph	Sentence	Film, Book, Sport, Music	5.8	1.0	91K
DuConv	Chinese	Text & Graph	Dialog	Film	9.1	2.0	270K
KdConv (ours)	**Chinese**	**Text & Graph**	**Sentence**	**Film, Music, Travel**	**19.0**	**2.3**	**86K**

Table 1: Comparison between our corpus and other human-labeled knowledge-grounded dialogue corpora.

- KdConv provides a benchmark to evaluate the ability of generating conversations with access to the corresponding knowledge in three domains. The corpus can empower the research of not only knowledge-grounded conversation generation, but also domain adaptation or transfer learning between similar domains (e.g., from film to music) or dissimilar domains (e.g., from music to travel).

- We provide benchmark models on this corpus to facilitate further research, and conduct extensive experiments. Results show that the models can be enhanced by introducing background knowledge, but there is still much room for further research. The corpus and the models are publicly available[3].

2 Related Work

Recently, open-domain conversation generation has been largely advanced due to the increase of publicly available dialogue data (Godfrey et al., 1992; Ritter et al., 2010; Shang et al., 2015; Lowe et al., 2015). However, the lack of annotation of background information or related knowledge results in significantly degenerated conversations, where the text is bland and strangely repetitive (Holtzman et al., 2019). These models produce conversations that are substantially different from those humans make, which largely rely on background knowledge.

To facilitate the development of conversational models that mimic human conversations, there have been several knowledge-grounded corpora proposed. Some datasets (Zhou et al., 2018b; Ghazvininejad et al., 2018; Liu et al., 2018; Tuan et al., 2019; Qin et al., 2019) collect dialogues and label the knowledge annotations using NER, string match, artificial scoring, and filtering rules based on external knowledge resources (Liu et al., 2018). However, mismatches between dialogues and knowledge resources introduce noises to these datasets. To obtain the high-quality knowledge-grounded datasets, some studies construct dialogues from scratch with human annotators, based on the unstructured text or structured knowledge graphs. For instance, several datasets (Zhou et al., 2018b; Dinan et al., 2018; Gopalakrishnan et al., 2019) have human conversations where one or both participants have access to the unstructured text of related background knowledge, while OpenDialKG (Moon et al., 2019) and DuConv (Wu et al., 2019) build up their corpora based on structured knowledge graphs. In Table 1, we present a survey on existing human-labeled knowledge-grounded dialogue datasets.

CMU DoG (Zhou et al., 2018b) utilizes 30 Wikipedia articles about popular movies as grounded documents, which explores two scenarios: only one participant has access to the document, or both have. Also using Wikipedia articles, however, Wizard of Wikipedia (WoW) (Dinan et al., 2018) covers much more dialogue topics (up to 1,365), which puts forward a high demand for the generalization ability of dialog generation models. One other difference from CMU DoG is that in WoW, only one participant has access to an information retrieval system that shows the worker paragraphs from Wikipedia possibly relevant to the conversation, which is unobservable to the other. In addition to the unstructured text, India DoG (Moghe et al., 2018) uses fact tables as background resources.

The idea of using structured knowledge to construct dialogue data is also adopted in OpenDialKG (Moon et al., 2019), which has a similar setting to KdConv. OpenDialKG contains chit-chat conversations between two agents engaging in a dialog about a given topic. It uses the Freebase knowl-

[3] https://github.com/thu-coai/KdConv

Domain	Film	Music	Travel	Total
# entities	7,477	4,441	1,154	13,072
# start	559	421	476	1,456
# extended	6,917	4,020	678	11,615
# relations	4,939	4,169	7	9,115
# triples	89,618	56,438	10,973	157,029
Avg. # triples per entity	12.0	12.7	9.5	12.0
Avg. # tokens per triple	20.5	19.2	20.9	20.1
Avg. # chars per triple	51.6	45.2	39.9	48.5

Table 2: Statistics of the knowledge graphs used in constructing KdConv (char represents character).

Domain	Film	Music	Travel	Total
# dialogues		1,500		4,500
# dialogues in Train/Dev/Test		1,200/150/150		3,600/450/450
# utterances	36,618	24,885	24,093	85,596
Avg. # utters per dialogue	24.4	16.6	16.1	19.0
Avg. # topics per dialogue	2.6	2.1	2.2	2.3
Avg. # tokens per utter	13.3	12.9	14.5	13.5
Avg. # characters per utter	20.4	19.5	22.9	20.8
Avg. # tokens per dialogue	323.9	214.7	233.5	257.4
Avg. # chars per dialogue	497.5	324.0	367.8	396.4
# entities	1,837	1,307	699	3,843
# start entities	559	421	476	1,456
# relations	318	331	7	656
# triples	11,875	5,747	5,287	22,909
Avg. # triples per dialogue	16.8	10.4	10.0	10.1
Avg. # tokens per triple	25.8	29.7	31.0	28.3
Avg. # chars per triple	49.4	56.8	57.4	53.6

Table 3: Statistics of KdConv (utter/char represents utterance/character respectively).

edge base (Bast et al., 2014) as background knowledge. In OpenDialKG, the entities and relations that are mentioned in the dialog are annotated, and it also covers multiple domains (film, books, sports, and music). However, the limitation is that there are much fewer turns in a conversation, and the whole dialogue is restricted to only one given topic, which is not suitable for modeling topic transition in human-like conversations.

To the best of our knowledge, DuConv (Wu et al., 2019) is the only existing Chinese human-labeled knowledge-grounded dialogue dataset. DuConv also utilizes unstructured text like short comments and synopsis, and structured knowledge graphs as knowledge resources. Given the knowledge graph, it samples two linked entities, one as the transitional topic and the other as the goal topic, to construct a conversation path. This path is used to guide participants toward the goal of the dialogue, which, as argued in Wu et al. (2019), can guide a model to deliver proactive conversations. However, the existence of the target path is inconsistent with an open dialogue in reality because humans usually do not make any assumption about the final topic of a conversation. Beyond that, the knowledge graph and the goal knowledge path are only annotated for the whole dialogue, which cannot provide explicit supervision on knowledge interactions for conversational models.

3 Dataset Collection

KdConv is designed to collect open-domain multi-turn conversations for modeling knowledge interactions in human-like dialogues, including knowledge planning, knowledge grounding, knowledge adaptations, etc. However, the open-domain background or commonsense knowledge is too large in scale (e.g., there are over 8 million concepts and 21 million relations in ConceptNet (Speer and Havasi, 2013)). Thus, it is costly and time-consuming to collect multi-turn conversations from scratch based on such large-scale knowledge. KdConv is proposed as one small step to achieve this goal, where we narrowed down the scale of background knowledge to several domains (film, music, and travel) and collected conversations based on the domain-specific knowledge. KdConv contains similar domains (film and music) and dissimilar domains (film and travel) so that it offers the possibility to investigate the generalization and transferability of knowledge-driven conversational models with transfer learning or meta learning(Gu et al., 2018; Mi et al., 2019).

In the following subsections, we will describe the two steps in data collection: (1) Constructing the domain-specific knowledge graph; (2) Collecting conversation utterances and knowledge interactions by crowdsourcing.

3.1 Knowledge Graph Construction

As the sparsity and the large scale of the knowledge were difficult to handle, we reduced the range of the domain-specific knowledge by crawling the most popular films and film stars, music and singers, and attractions as start entities, from several related websites for the film[4]/music[5]/travel[6] domain. The knowledge of these start entities contains both structured knowledge triples and unstructured knowledge texts, which make the task more general but challenging. After filtering the start entities which have few knowledge triples, the film/music/travel domain contains 559/421/476

[4]https://movie.douban.com/top250
[5]https://music.douban.com/top250
[6]https://travel.qunar.com/p-cs299914-beijing-jingdian

start entities, respectively.

After crawling and filtering the start entities, we built the knowledge graph for each domain. Given the start entities as seed, we retrieved their neighbor entities within three hops from XLORE, a large-scale English-Chinese bilingual knowledge graph (Wang et al., 2013). We merged the start entities and these retrieved entities (nodes in the graph) and relations (edges in the graph) into a domain-specific knowledge graph for film and music domains. For the travel domain, we built the knowledge graph with the knowledge crawled only from the Web, because XLORE provides little knowledge for start entities in the travel domain. There are two types of entities in the knowledge graph: one is the start entities crawled from the websites, the other is the extended entities that are retrieved from XLORE (film/music), or websites (travel) to provide related background knowledge. The statistics of the knowledge graphs used in constructing KdConv are provided in Table 2.

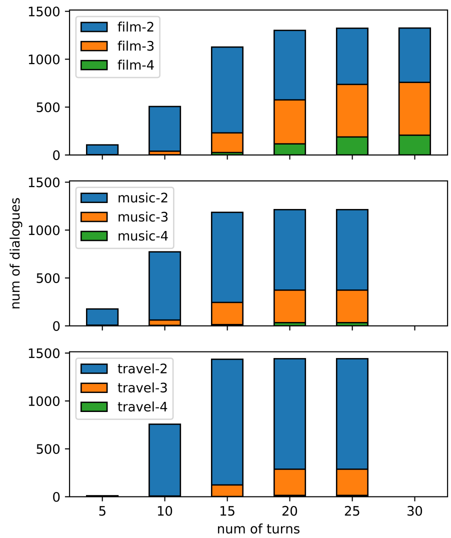

Figure 2: Statistics of the number of dialogues where at least $k (k = 2, 3, 4)$ topics have been discussed in the first n turns. The proportions of dialogues that contain 3 or 4 topics become larger when the dialog turn becomes longer.

3.2 Dialogue Collection

We recruited crowdsourced annotators to generate multi-turn conversations that are related to the domain-specific knowledge graph without any predefined goals or constraints. During the conversation, two speakers both had access to the knowledge graph rather than that only one participant had access to the knowledge, as proposed in WoW (Dinan et al., 2018) where one party always leads the conversation with an expert-apprentice mode. Allowing two participants to access the knowledge, in our corpus the two parties can dynamically change their roles, as either leader or follower, which is more natural and real to human conversations. In addition to making dialogue utterances, the annotators were also required to record the related knowledge triples if they generated an utterance according to some triples. To increase the knowledge exposure in the collected conversations, the annotators were instructed to start the conversation based on one of the start entities, and they were also encouraged to shift the topic of the conversation to other entities in the knowledge graph. Thus, the topics of conversations and the knowledge interactions in KdConv are diversified and unconstrained. In order to ensure the naturalness of the generated conversations, we filtered out low-quality dialogues, which contain grammatical errors, inconsistencies of knowledge facts, etc. The distinct-4 score is 0.54/0.51/0.42 for the film/music/travel domain, which is comparable to the score of DuConv (Wu et al., 2019), 0.46. The distinct-4 score decreases, due to the decrease of knowledge triples and utterances in three domains, as shown in Table 3.

3.3 Corpus Statistics

The detailed statistics of KdConv are shown in Table 3. We collect 1,500 dialogues for each domain. The training, validation, and test sets are partitioned with the ratio of 8:1:1. Note that the number of conversation turns in the film domain is larger than those in the music/travel domains (24.4 vs. 16.6/16.1), while the utterance lengths are similar (13.3 vs. 12.9/14.5 at the token level, and 20.4 vs. 19.5/22.9 at character level). As aforementioned, the dialogues in the real world are not limited to one or two topics, while discussing multiple topics in depth usually requires a conversation having enough number of turns. In order to verify this point, we analyze the relationship between the number of turns and the number of topics. Note that the topics are defined as the distinct head entities in the knowledge triples and the central nodes with a degree greater than 1 in the knowledge graph.

The results of three domains are shown in Figure 2. Given a number $k (k = 2, 3, 4)$ of topics and a number n of conversation turns, we count the number of dialogues where at least k topics have been discussed in the first n turns. It can be observed that more topics tend to appear in a dialogue only if there are enough conversation turns. For instance, most dialogues involve at least 2 topics when the number of turns exceeds 15. This is consistent with the fact that if a conversation is very short, speakers will not be able to discuss in detail, let alone natural transition between multiple topics.

	Topic Transition
1 Hop	T_1 −Major Work→ T_2 T_1 −Star→ T_2 T_1 −Director→ T_2
2 Hop	T_1 −Major Work→ T_2 −Star→ T_3 T_1 −Major Work→ T_2 −Director→ T_3 T_1 −Star→ T_2 −Major Work→ T_3
3 Hop	T_1 −Major Work→ T_2 −Star→ T_3 −Major Work→ T_4 T_1 −Star→ T_2 −Major Work→ T_3 −Director→ T_4 T_1 −Major Work→ T_2 −Star→ T_3 −Information→ T_4

Table 4: Top-3 topic transition of the film domain, where T_n denotes the n-th topic of a dialog and $T_n - X \rightarrow T_{n+1}$ represents the relation X between T_n and T_{n+1}.

To analyze topic transition in our dataset, we provide top-3 topic transition in the film domain, as shown in Table 4. As can be seen, topic transition has diverse patterns conditioned on different hops. With the increase of the hops of topic transition, the complexity of topic transition goes up. Compared to DuConv (Wu et al., 2019), the dialogues of KdConv contain multiple and diverse topics instead of fixed two topics, leading to diverse and complex topic transition, which are more suitable for the research of knowledge planning in human-like conversations. Note that the relation "−Information→" appeared in the last row is different from the other relations, which means the target topic is mentioned in unstructured texts describing the information about the source topic. The low frequency of the relation "−Information→" demonstrates that people prefer to shift the topic according to the structured relations rather than unstructured texts, as adopted in WoW (Dinan et al., 2018).

4 Experiments

4.1 Models

To provide benchmark models for knowledge-driven conversation modeling, we evaluated both generation- and retrieval-based models on our corpus. In order to explore the role of knowledge annotation, we evaluated the models with/without access to the knowledge graph of our dataset.

4.1.1 Generation-based Models

Language Model (LM) (Bengio et al., 2003): We trained a language model that maximizes the log likelihood: $\log \mathcal{P}(\boldsymbol{x}) = \sum_t \log \mathcal{P}(x_t|x_{<t})$, where $\boldsymbol{x}$ denotes a long sentence that sequentially concatenates all the utterances of a dialogue.

Seq2Seq (Sutskever et al., 2014): An encoder-decoder model augmented with attention mechanism (Bahdanau et al., 2014). The input of the encoder was the concatenation of the past $k-1$ utterances, while the target output of the decoder was the k-th utterance. k was set to 8 in the experiment. If there were fewer than $k-1$ sentences in the dialogue history, all the past utterances would be used as input.

HRED (Serban et al., 2016): A hierarchical recurrent encoder-decoder model that has a specific context RNN to incorporate historical conversational utterances into a context state, which is used as the initial hidden state of the decoder. The adapted model generates the k-th utterance based on the past $k-1$ utterances, where k was also set to 8, for fair comparison with Seq2Seq.

All the generative models were trained by optimizing the cross-entropy loss:

$$\mathcal{L}_0^{(g)} = -\frac{1}{T}\sum_{t=1}^{T} \log \mathcal{P}(\hat{x}_t = x_t),$$

where $\hat{x}_t$ denotes the predicted token at the time step t, while x_t is the t-th token of the target sentence.

4.1.2 Retrieval-based Model

BERT (Devlin et al., 2019): We adapted this deep bidirectional transformers (Vaswani et al., 2017) as a retrieval-based model. For each utterance (except the first one in a dialog), we extracted keywords in the same way as Wu et al. (2017) and retrieved 10 response candidates, including the golden truth based on the BM25 algorithm (Robertson et al., 1995). The training task is to predict whether a candidate is the correct next utterance given the context, where a sigmoid function was used to output the probability score $\hat{y} = \mathcal{P}(y = 1)$ and the cross-entropy loss was optimized:

$$\mathcal{L}_0^{(r)} = -y \log \hat{y} - (1-y)\log(1-\hat{y}),$$

where $y \in \{0, 1\}$ is the true label. For the test, we selected the candidate response with the largest probability.

4.1.3 Knowledge-aware Models

A key-value memory module (Miller et al., 2016) is introduced to the aforementioned models to utilize the knowledge information. We treated all knowledge triples mentioned in a dialogue as the knowledge information in the memory module. For a triple that is indexed by i, we represented the key memory and the value memory respectively as a key vector $\boldsymbol{k}_i$ and a value vector $\boldsymbol{v}_i$, where $\boldsymbol{k}_i$ is the average word embeddings of the head entity and the relation, and $\boldsymbol{v}_i$ is those of the tail entity. We used a query vector $\boldsymbol{q}$ to attend to the key vectors $\boldsymbol{k}_i (i = 1, 2, ...)$: $\alpha_i = \text{softmax}_i(\boldsymbol{q}^T \boldsymbol{k}_i)$, then the weighted sum of the value vectors $\boldsymbol{v}_i (i = 1, 2, ...)$, $\boldsymbol{v} = \sum_i \alpha_i \boldsymbol{v}_i$, was incorporated into the decoding process (for the generation-based models, concatenated with the initial state of the decoder) or the classification (for the retrieval-based model, concatenated with the <CLS> vector). For Seq2Seq, $\boldsymbol{q}$ is the final hidden state of the encoder. For HRED, we treated the context vector as the query, while for BERT, the output vector of <CLS> was used.

Note that our dataset has a sentence-level annotation on the knowledge triples that each utterance uses. To force the knowledge-aware models to attend to the golden KG triples, we added an extra attention loss (for retrieval-based models, this loss was computed only on the positive examples):

$$\mathcal{L}_{\text{att}} = -\frac{1}{|\{\text{truth}\}|} \sum_{i \in \{\text{truth}\}} \log \alpha_i,$$

where $\{\text{truth}\}$ is the set of indexes of triples that are used in the true response. The total loss are the weighted sum of $\mathcal{L}_0^{(l)}$ and $\mathcal{L}_{\text{att}}$:

$$\mathcal{L}_{\text{tot}}^{(l)} = \mathcal{L}_0^{(l)} + \lambda \mathcal{L}_{\text{att}}, \quad l \in \{g, r\}.$$

Note that the knowledge-enhanced BERT was initialized from the fine-tuned BERT discussed in Section 4.1.2, and the parameters of the transformers were frozen during training the knowledge related modules. The purpose was to exclude the impact of the deep transformers but only examine the potential effects introduced by the background knowledge.

4.2 Implementation Details

We implemented the above models with TensorFlow (Abadi et al., 2016), PyTorch (Paszke et al., 2017) and CoTK (Huang et al., 2020). The Jieba Chinese word segmenter[7] was employed for tokenization. The 200-dimensional word embeddings were initialized by Song et al. (2018), while the unmatched ones were randomly sampled from a standard normal distribution $\mathcal{N}(0, 1)$. The type of RNN network units was all GRU (Cho et al., 2014) and the number of hidden units of GRU cells were all set to 200. ADAM (Kingma and Ba, 2014) was used to optimize all the models with the initial learning rate of 5×10^{-5} for BERT and 10^{-3} for others. The mini-batch sizes are set to 2 dialogues for LM and 32 pairs of post and response for Seq2Seq and HRED.

4.3 Automatic Evaluation

4.3.1 Metrics

We measured the performance of all the retrieval-based models using Hits@1 and Hits@3, same as Zhang et al. (2018) and Wu et al. (2019). [8] We adopted several widely-used metrics to measure the quality of the generated response. We calculated Perplexity (PPL) to evaluate whether the generation result is grammatical and fluent. BLEU-1/2/3/4 (Papineni et al., 2002) is a popular metric to compute the k-gram overlap between a generated sentence and a reference (Sordoni et al., 2015; Li et al., 2016b). Distinct-1/2/3/4 (Li et al., 2016b) is also provided to evaluates the diversity of generated responses.

4.3.2 Results

The results are shown in Table 5. We analyze the results from the following perspectives:

The influence of knowledge: after introducing the knowledge, all the models were improved in terms of all the metrics except PPL in all the domains. First, all the models obtain higher Hits@1 scores (in the music domain, BERT obtains an improvement of 0.4 on Hits@1). After incorporating the knowledge into BERT, the performance of Hits@1 improves slightly, because the memory network which models knowledge information is rather shallow, compared to the deep structure in BERT. Second, Seq2Seq and HRED both have better BLEU-k scores (in the travel domain, Seq2Seq obtains an improvement of 7.2 on BLEU-4), which means a better quality of generated responses. Third, the two generation-based models

[7] https://github.com/fxsjy/jieba

[8] For generative models, the rank is decided by the PPL values of candidate responses.

Model	Hits@1/3		PPL	BLEU-1/2/3/4				Distinct-1/2/3/4			
				Film							
LM	14.30	35.70	**21.91**	24.22	12.40	7.71	4.27	2.32	6.13	10.88	16.14
Seq2Seq	17.54	40.57	23.88	26.97	14.31	8.53	5.30	2.51	7.14	13.62	21.02
HRED	16.45	40.62	24.74	27.03	14.07	8.30	5.07	2.55	7.35	14.12	21.86
BERT	65.36	<u>91.79</u>	-	81.64	77.68	75.47	73.99	8.55	31.28	51.29	63.38
Seq2Seq + know	**17.77**	**41.66**	25.56	27.45	14.51	8.66	5.32	2.85	7.98	15.09	23.17
HRED + know	17.38	39.79	26.27	**27.94**	**14.69**	**8.73**	**5.40**	**2.86**	**8.08**	**15.81**	**24.93**
BERT + know	<u>65.67</u>	<u>91.79</u>	-	<u>81.98</u>	<u>78.08</u>	<u>75.90</u>	<u>74.44</u>	<u>8.59</u>	<u>31.47</u>	<u>51.63</u>	<u>63.78</u>
				Music							
LM	18.09	39.36	**14.61**	25.80	13.93	8.61	5.57	2.72	7.31	12.69	18.64
Seq2Seq	22.65	44.43	16.17	28.89	16.56	10.63	7.13	2.52	7.02	12.69	18.78
HRED	21.20	42.84	16.82	**29.92**	17.31	11.17	7.52	2.71	7.71	14.07	20.97
BERT	55.64	<u>86.90</u>	-	78.71	73.61	70.55	68.43	6.57	26.75	44.75	55.85
Seq2Seq + know	**22.90**	**47.14**	17.12	29.60	17.26	11.36	7.84	**3.93**	**12.35**	**23.01**	**34.23**
HRED + know	21.82	45.33	17.69	29.73	**17.51**	**11.59**	**8.04**	3.80	11.70	22.00	33.37
BERT + know	<u>56.08</u>	86.87	-	<u>78.98</u>	<u>73.91</u>	<u>70.87</u>	<u>68.76</u>	<u>6.59</u>	<u>26.81</u>	<u>44.84</u>	<u>55.96</u>
				Travel							
LM	22.16	41.27	**8.86**	27.51	17.79	12.85	9.86	3.18	8.49	13.99	19.91
Seq2Seq	27.07	46.34	10.44	29.61	20.04	14.91	11.74	3.75	11.15	19.01	27.16
HRED	25.76	46.11	10.90	30.92	20.97	15.61	12.30	4.15	12.01	20.52	28.74
BERT	45.25	71.87	-	81.12	76.97	74.47	72.73	7.17	22.55	34.03	40.78
Seq2Seq + know	**29.67**	**50.24**	10.62	**37.04**	**27.28**	**22.16**	**18.94**	**4.25**	**13.64**	**24.18**	34.08
HRED + know	28.84	49.27	11.15	36.87	26.68	21.31	17.96	3.98	13.31	24.06	**34.35**
BERT + know	<u>45.74</u>	<u>71.91</u>	-	<u>81.28</u>	<u>77.17</u>	<u>74.69</u>	<u>72.97</u>	<u>7.20</u>	<u>22.62</u>	<u>34.11</u>	<u>40.86</u>

Table 5: Automatic evaluation. The best results of generative models and retrieval models are in **bold** and <u>underlined</u> respectively. "+ know" means the models enhanced by the knowledge base.

also gain larger Distinct-k values (in the music domain, HRED obtains an improvement of 12.4 on Distinct-4), which indicates a better diversity of the generated results.

Comparison between models: In all the three domains, the knowledge-aware BERT model achieves the best performance in most of the metrics, as it retrieves the golden-truth response at a fairly high rate. HRED performs best in BLEU-k and Distinct-k among all the generation-based baselines without considering the knowledge. Knowledge-aware HRED has better results of BLEU-k in the film and music domains and better results of Distinct-k in the film domain, while the knowledge-enhanced Seq2Seq achieves the best Hits@1/3 scores among all the generation-based models.

Comparison between domains: For retrieval-based models, the performance is best in the film domain but worst in the travel domain, largely affected by the data size (see Table 3). For generation-based models, however, the performance improves from the film domain to the travel domain, as the average number of utterances per dialogue decreases from 24.4 in the film domain to 16.1 in the travel domain (see Table 3). The more utterances a dialogue contains, the more difficulties in conversation modeling for generation-based models. Besides, the more diverse knowledge (1,837 entities and 318 relations in the film domain, vs. 699 entities and 7 relations in the travel domain) also requires the models to leverage knowledge more flexibly. The difference between different domains can be further explored in the setting of transfer learning or meta learning in the following research.

4.4 Manual Evaluation

To better understand the quality of the generated responses from the semantic and knowledge perspective, we conducted the manual evaluation for knowledge-aware BERT, knowledge-aware HRED, and HRED, which have achieved advantageous performance in automatic evaluation[9].

4.4.1 Metrics

Human annotators were asked to score a generated response in terms of the fluency and coherence

[9]We omitted the BERT model because it performs similarly to knowledge-aware BERT as shown in automatic evaluation.

Model	Fluency	Coherence
Film \ κ	0.50	0.61
HRED	1.64	1.19
HRED + know	*1.78*	*1.28*
BERT + know	**2.00**	**1.79**
Music \ κ	0.37	0.57
HRED	1.90	1.30
HRED + know	1.86	1.36
BERT + know	**2.00**	**1.80**
Travel \ κ	0.55	0.74
HRED	1.77	1.10
HRED + know	1.78	*1.31*
BERT + know	**2.00**	**1.76**

Table 6: Manual evaluation. The best results (t-test, p-value < 0.005) are in **bold**. Between two generative models, the significantly better results are *italic underlined* (t-test, p-value < 0.005) or underlined (t-test, p-value < 0.05). κ is the Fleiss' kappa value. "+ know" means the models enhanced by knowledge information.

metrics. The fluency score (rating scale is 0,1,2) is defined as whether the response is fluent and natural. The coherence (rating scale is 0,1,2) is defined as whether a response is relevant and coherent to the context and the knowledge information.

4.4.2 Annotation Statistics

We randomly sampled about 500 contexts from the test sets of the three domains and generated responses by each model. These 1,500 context-response pairs in total and related knowledge graphs were presented to three human annotators.

We calculated the Fleiss' kappa (Fleiss, 1971) to measure inter-rater consistency. Fleiss' kappa for Fluency and Coherence is from 0.37 to 0.74, respectively. The overall 3/3[10] agreement for Fluency and Coherence is from 68.14% to 81.33% in the three domains.

4.4.3 Results

The results are shown in Table 6. As can be seen, knowledge-aware BERT outperforms other models significantly in both metrics in all the three domains, which agrees with the results of automatic evaluation. The Fluency is 2.00 because the retrieved responses are all human-written sentences. The Fluency scores of both generation-based models are close to 2.00 (in the music domain, the Fluency of HRED is 1.90), showing that the generated responses are fluent and grammatical. The Coherence scores of both HRED and knowledge-aware HRED are higher than 1.00 but still have a huge gap to 2.00, indicating that the generated responses are relevant to the context but not coherent to knowledge information in most cases. After incorporating the knowledge information into HRED, the Coherence score is improved significantly in all the three domains, as the knowledge information is more expressed in the generated responses.

5 Conclusion and Future Work

In this paper, we propose a Chinese multi-domain corpus for knowledge-driven conversation generation, KdConv. It contains 86K utterances and 4.5K dialogues, with an average number of 19.0 turns. Each dialogue contains various topics and sentence-level annotations that map each utterance with the related knowledge triples. The dataset provides a benchmark to evaluate the ability to model knowledge-driven conversations. In addition, KdConv covers three domains, including film, music, and travel, that can be used to explore domain adaptation or transfer learning for further research. We provide generation- and retrieval-based benchmark models to facilitate further research. Extensive experiments demonstrate that these models can be enhanced by introducing knowledge, whereas there is still much room in knowledge-grounded conversation modeling for future work.

Acknowledgments

This work was jointly supported by the NSFC projects (Key project with No. 61936010 and regular project with No. 61876096), and the National Key R&D Program of China (Grant No. 2018YFC0830200). We thank THUNUS NExT Joint-Lab for the support.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly

[10] 3/3 means all the three annotators assign the same label to an annotation item.

learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Hannah Bast, Florian Bäurle, Björn Buchhold, and Elmar Haußmann. 2014. Easy access to the freebase dataset. In *WWW*, pages 95–98. ACM.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

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

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*.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520. IEEE.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-TÃijr. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Jiatao Gu, Yong Wang, Yun Chen, Victor OK Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *EMNLP*, pages 3622–3631.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Fei Huang, Dazhen Wan, Zhihong Shao, Pei Ke, Jian Guan, Yilin Niu, Xiaoyan Zhu, and Minlie Huang. 2020. Cotk: An open-source toolkit for fast development and fair evaluation of text generation. *arXiv preprint arXiv:2002.00583*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016b. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 1489–1498.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Fei Mi, Minlie Huang, Jiyong Zhang, and Boi Faltings. 2019. Meta-learning for low-resource natural language generation in task-oriented dialogue systems. In *IJCAI*.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *EMNLP*, pages 1400–1409.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *EMNLP*, pages 2322–2332.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *ACL*, pages 845–854.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by reading: Contentful neural conversation with on-demand machine reading. *arXiv preprint arXiv:1906.02738*.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *NAACL*, pages 172–180.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, pages 3776–3784.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *ACL*, pages 1577–1586.

Louis Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating long and diverse responses with neural conversation models. *arXiv preprint arXiv:1701.03185*.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *NAACL*, pages 175–180.

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

Robert Speer and Catherine Havasi. 2013. Conceptnet 5: A large semantic network for relational knowledge. In *The PeopleâĂŹs Web Meets NLP*, pages 161–176. Springer.

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

Yi-Lin Tuan, Yun-Nung Chen, and Hung-yi Lee. 2019. Dykgchat: Benchmarking dialogue generation grounding on dynamic knowledge 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 1855–1865.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Zhigang Wang, Juanzi Li, Zhichun Wang, Shuangjie Li, Mingyang Li, Dongsheng Zhang, Yao Shi, Yongbin Liu, Peng Zhang, and Jie Tang. 2013. Xlore: A large-scale english-chinese bilingual knowledge graph. In *International semantic web conference (Posters & Demos)*, volume 1035, pages 121–124.

Tsung Hsien Wen, Milica Gasic, Nikola Mrksic, Pei Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *EMNLP*, pages 1711–1721.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804, Florence, Italy. Association for Computational Linguistics.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Vancouver, Canada. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Hao Zhou, Minlie Huang, and Xiaoyan Zhu. 2016. Context-aware natural language generation for spoken dialogue systems. In *COLING*, pages 2032–2041.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018a. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018b. A dataset for document grounded conversations. In *EMNLP*, pages 708–713.

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