Augmenting Knowledge-grounded Conversations with Sequential Knowledge Transition

Haolan Zhan[1]*, **Hainan Zhang**[2]†, **Hongshen Chen**[2], **Zhuoye Ding**[2]
Yongjun Bao[2] and **Yanyan Lan**[3]
[1] Institute of Software, Chinese Academy of Sciences, Beijing, China
[2]Data Science Lab, JD.com, Beijing, China
[3]Institute for AI Industry Research, Tsinghua University, Beijing, China
zhanhaolan316@gmail.com, zhanghainan6@jd.com, ac@chenhongshen.com,
dingzhuoye@jd.com, lanyanyan@tsinghua.edu.cn

Abstract

Knowledge data are massive and widespread in the real-world, which can serve as good external sources to enrich conversations. However, in knowledge-grounded conversations, current models still lack the fine-grained control over knowledge selection and integration with dialogues, which finally leads to the knowledge-irrelevant response generation problems: 1) knowledge selection merely relies on the dialogue context, ignoring the inherent knowledge transitions along with conversation flows; 2) the models often over-fit during training, resulting with incoherent response by referring to unrelated tokens from specific knowledge content in the testing phase; 3) although response is generated upon the dialogue history and knowledge, the models often tend to overlook the selected knowledge, and hence generates knowledge-irrelevant response. To address these problems, we proposed to explicitly model the knowledge transition in sequential multi-turn conversations by abstracting knowledge into topic tags. Besides, to fully utilizing the selected knowledge in generative process, we propose pre-training a knowledge-aware response generator to pay more attention on the selected knowledge. In particular, a sequential knowledge transition model equipped with a pre-trained knowledge-aware response generator (SKT-KG) formulates the high-level knowledge transition and fully utilizes the limited knowledge data. Experimental results on both structured and unstructured knowledge-grounded dialogue benchmarks indicate that our model achieves better performance over baseline models.

1 Introduction

Knowledge-grounded conversations (Long et al., 2017; Liu et al., 2018; Niu et al., 2019; Xu et al., 2020), aiming at improving the informativeness and specificity of dialogue generation by exploiting external knowledge sources, has attracted much attention as a potential solution to relieve the common response problem (Li et al., 2015; Zhang et al., 2018a; Ren et al., 2020) in dialogue generation, i.e.,*'I don't know.'* and *'What do you mean?'*. Typically, knowledge-grounded conversation is decomposed into two sub-processes (Dinan et al., 2018; Wu et al., 2019): knowledge selection (KS) based on dialogue context, and response generation with reference to the selected knowledge. Therefore, to select relevant knowledge and then incorporate it efficiently, is of great significance for multi-turn knowledge-grounded dialogue generation task.

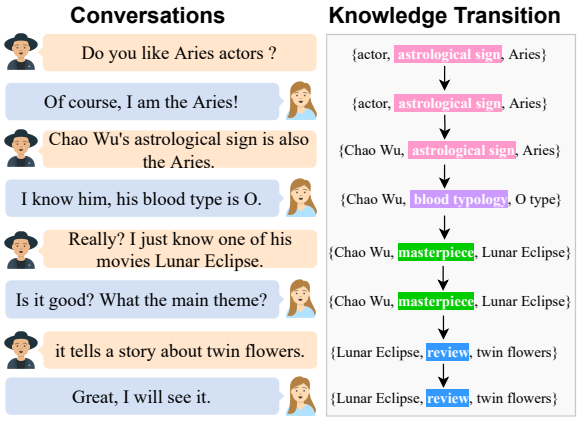

Figure 1: The example from the DuConv dataset (Wu et al., 2019), shows the knowledge transition in real dialogue.

Although external knowledge sources are widespread in the real-world, in fact, current knowledge-grounded conversations still lack the fine-grained control over knowledge selection and integration with dialogues. Most existing works (Liu et al., 2018; Niu et al., 2019) select knowledge according to the given dialogue context (Lian et al., 2019; Kim et al., 2020). However, the sequential transition characteristic of knowledge (also known as knowledge shift) along multiple sequential conversation turns is neglected. As

*Work done at Data Science Lab, JD.com.
†Corresponding author.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5621–5630
June 6–11, 2021.

shown in Figure 1, two people are talking about an actor from the knowledge "*astrological sign*" to another knowledge "*blood typology*", which is a natural transition in human personality chat (Mayo et al., 1978; Miller, 2014). By nature, taking the knowledge sequential transition characteristic into account is of tremendous benefits to the knowledge grounded conversations.

What's more, knowledge-irrelevant response generation problem also hampers the performance of existing models. This is caused by two reasons. The first reason is that current models often over-fit during training, resulting with incoherent response by referring to unrelated tokens from specific knowledge content in testing phase. To resolve this problem, we propose to calculate the knowledge transition probability among different turns on a high-level representation, i.e., knowledge topic tag. With such concise high-level knowledge representation, our model is not limited to conventional structured knowledge-grounded conversation but can be easily adapted to unstructured knowledge-based conversations. For example, in structured triple data, i.e.,{obj, relation, content}, we can utilize the "relation" as the high-level topic tag to model the sequential knowledge transition process in conversations. As shown in Figure 1, the topic migrates from the "*astrological sign* ' " tag to the "*blood typology*" tag, and then moves to the "*masterpiece*". In the unstructured dataset like 'Wizard of Wikipedia' (Dinan et al., 2018), we can utilize topic models, such as LDA (Blei et al., 2003), to obtain the knowledge tag for each turn, and then calculate the sequential transition probability among these tags. Since the number of tag categories is limited, it can be well employed to model the knowledge transition.

Moreover, the second reason is that the models often tends to overlook the selected knowledge, and hence generates knowledge-irrelevant response. To address this problem, we propose pre-training a knowledge-aware response generator, aiming at generating a natural sentence based on a given knowledge, in order to make full use of the limited knowledge data. For example in Figure 1, given the triple *' {Chao Wu, astrological sign, Aries}'* , the knowledge-aware generator is optimized to generate a sentence *' Chao Wu's astrological sign is Aries.'*. Obviously, the generator should also has the ability to generate *' Zhiling Lin's astrological sign is Virgo.'* while given *'{Zhiling Lin, astrological sign, Virgo}'*. Actually, the knowledge-aware response generator learns how to generate a natural sentence based on a relation tag rather than the knowledge content. It is like that one student learns grammar rules rather than specific examples while learning a foreign language. Therefore, even with the limited data, the generator can also generate relevant sentences about given knowledge.

In this paper, we propose a sequential knowledge transition model equipped with a pre-trained knowledge-aware response generator (SKT-KG), which can conduct the high-level knowledge transition in conversation and fully use of the limited knowledge data. Specifically, at first, we pre-train a transformer-based response generator based on the knowledge. And then, we utilize a BiLSTM-CRF (Huang et al., 2015) network to model the knowledge transition process, and select the knowledge tag with maximum score and its corresponding knowledge content. Finally, we feed the dialogue utterances and the selected knowledge content together into the pre-trained knowledge-aware response generator to generate final response.

In our experiments, we use two public knowledge-grounded dialogue datasets to evaluate our proposed models, i.e. structured DuConv corpus and unstructured Wizard of Wikipedia (WoW) corpus. The results show that our SKT-KG model has the ability to produce more diverse and suitable responses than traditional knowledge-grounded models. Besides, we conduct an analysis on knowledge selection, and the results show that the SKT-KG model obtains higher ranking measure than baselines, which indicates that the knowledge selected by our model is reasonable.

2 Related Work

Recently, dialogue systems have gained more attention in both research community (Vougiouklis et al., 2016; Liu et al., 2018; Zhou et al., 2018; Shen et al., 2019; Shen and Feng, 2020) and industry (Xu et al., 2020; Zhao et al., 2020), because of its practicality in the real application, such as chatbot and customer services (Chen et al., 2020; Liu et al., 2020; Shen et al., 2021; Zhang et al., 2019; Chen et al., 2018; Zhang et al., 2020). With external knowledge sources, dialogue systems can generate more specific and informative response, which has great potential to resolve the common response problem (Zhang et al., 2018b; Ren et al., 2020). The majority of previous works decomposed the

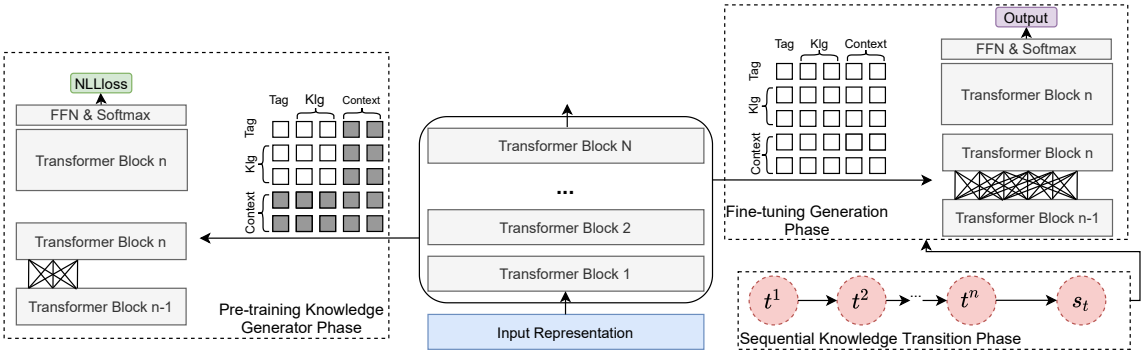

Figure 2: The architecture of SKT-KG model. The left shows the pre-training phase of knowledge-aware response generator with a flexible self-attention mask mechanism. The bottom-right shows the knowledge transition module, which can select the knowledge sequentially along with conversations. And the top-right shows the fine-tuning phase to generate a response based on the selected knowledge and dialogue utterances in history.

knowledge-grounded dialogue generation task into two sub-problems: knowledge selection and response selection.

In knowledge selection, previous works proposed to use the keyword matching (Ghazvininejad et al., 2018; Liu et al., 2018), information retrieval (Young et al., 2018) and entity diffusion (Liu et al., 2018) methods to detect the relevant knowledge based on dialogue context, and finally feed both dialogue utterances and the selected knowledge into generative models. Specifically, Zhou et al. (2018) proposed to employ the graph attention mechanism to encode the retrieved relevant knowledge graph, which can augment the semantic understanding of dialogue context. Lian et al. (2019) proposed to use the prior and posterior distributions over knowledge to facilitate knowledge selection. Although these work are capable to model the relationship between context and knowledge, they still ignored the knowledge transition characteristic, which is important for knowledge selection.

Human dialogue depends on both local information and global information. Peng et al. (2019) also pointed out that natural language understanding requires a coherent understanding of a series of events or actions, not only what events have appeared, but also what is likely to happen next. Therefore, it is critical to obtain the natural and relevant knowledge for the knowledge-grounded dialogue generation. Sun et al. (2020) proposed to recurrently update the knowledge based on conversation history and progressively incorporate it into the history step-by-step. But they only consider the relationship of history to knowledge. However, these models may also suffer from a knowledge sparse problem, due to the low-resource limitation in reality (Zhao et al., 2020).

In reality, sufficient knowledge-grounded dialogues data are difficult to obtain. To tackle this practical challenge, Su et al. (2020) proposed to augment the dialogue generation with external non-conversational text, which may also introduce much noise. Li et al. (2020) proposed to pre-train the knowledge encoder with unstructured knowledge and fine-tune the model using the limited knowledge-grounded training examples. In our work, we propose to make full use of our training data and model the high-level knowledge transition process, which can resolve the sparse problem in knowledge-grounded dialogue data.

3 Approach

In this section, we propose a novel sequential knowledge transition model with pre-trained knowledge-aware response generator (SKT-KG), as shown in Figure 2. This model contains three major parts: pre-trained knowledge-aware response generator, sequential knowledge transition, and transformer decoder. Specifically, we firstly pre-train a transformer-based knowledge-aware response generator based on the knowledge and its corresponding natural sentence. And then, we utilize a BiLSTM-CRF (Huang et al., 2015) network to model the knowledge transition process, and select knowledge tag with maximum score and its corresponding knowledge content. Finally, we feed the context utterances and this selected knowledge content into the knowledge-aware response generator to fine-tune it. After fine-tuning, response can be generated by given the selected knowledge tag and corresponding content, and history dialogue utterances.

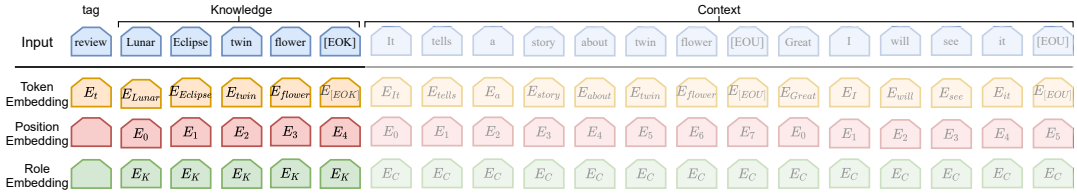

Figure 3: An example for the input representation. In the pre-training phase, we mask the context utterances part. And for the fine-tune response generation phase, we concatenate the selected knowledge tag, the selected knowledge content and history utterances as input.

3.1 Input Representation

Firstly, we introduce the data formulation in our model. Given the history knowledge content $K = \{k^1, \cdots, k^n\}$, the history context $C = \{c^1, \cdots, c^n\}$ and the candidate knowledge set for response $CK = \{ck^1, \cdots, ck^m\}$, the goal of our model is to select the most relevant and natural knowledge $ck^t \in CK$ based on the sequential K and C, and then generate the response $Y = \{y_1, \cdots, y_{|L|}\}$ based on the selected knowledge ck^t and context C. It is worth noting that each history utterance c^i is related to a history knowledge k^i and each knowledge k^i has a knowledge tag $t^i \in T$, which is explicit in the structured knowledge, such as ' relation' in triple knowledge as shown in Figure 1, and implicit in the unstructured knowledge, which is abstracted by topic model, i.e., LDA (Blei et al., 2003). Knowledge tag category $T = \{t^1, ..., t^N\}$ has N different knowledge tags.

We utilize the classical transformer blocks as the backbone framework. To generate response Y, the original input is the concatenation of the selected knowledge tag s_t, the selected knowledge content ck^t and the history context utterances$\{c^1, \cdots, c^n\}$. We use three different embedding methods for the original input: Token embedding, Role embedding and Position embedding, as shown in Figure 3. For knowledge content and dialogue utterances, we utilize the word embedding of each token as the token embedding. For knowledge tag, we map each tag to different categories as the token embedding. A special end-of-knowledge [EOK] token is inserted between knowledge and utterance context to mark the border. Another token end-of-utterance [EOU] is added at the end of each history dialog utterance. Role embeddings are employed to differentiate knowledge content and dialogue utterances. The role embedding E_K is added for the knowledge content, as well as dialogue utterances are represented by role embedding E_C. Position embeddings are added according to the token position in each utterance. Note that for the special token of knowledge tag, its corresponding role and position embeddings are both set to zero.

3.2 Pre-trained Knowledge-aware Response Generator

In our pre-trained knowledge-aware response generator, there are two essential phases we should consider: pre-training phase and fine-tuning response generation phase. In the pre-training phase, given the knowledge tag and knowledge content, our generator focuses on generating the relevant sentence, as shown in the left of Figure 2. And in the fine-tuning response generation phase, given the context utterances, the knowledge tag and the selected knowledge content, our generator focuses on generating the natural and relevant response, as shown in the top-right of Figure 2. To unify the pre-training phase and fine-tuning phase, we propose to utilize the flexible self-attention mask mechanism to distinguish the input representation in this two phases, as shown in Figure 3.

In the pre-training phase, we employ a self-attention mask mechanism to the history dialogue utterances, in order to train the knowledge-aware response generator independently. Given the knowledge content $k^i \in K$, its knowledge tag $t^i \in T$ and its corresponding sentence $c^i = \{x_1^i, \cdots, x_N^i\}$, we choose the negative log-likelihood loss as our training optimization.

$$\mathcal{L}_{pre}(\theta) = -\sum_{t=1}^{N} \log p(x_t^i | x_{<t}^i, k^i, t^i; \theta),$$

where θ denotes the model parameters and $x_{<t}^i$ denotes the previously generated words.

3.3 Sequential Knowledge Transition

In this section, we will introduce the knowledge selection process, including the utterance encoding and transition modules. To obtain the next knowledge tag, we should consider both the sequential

knowledge tags and the sequential context utterances, as shown in Figure 4.

Utterance Encoding. To conduct the context sequential representation, we use the standard base BERT model with average pooling (Cer et al., 2018) and the BiLSTM to obtain the context sequential representation. Given the context utterances $C = \{c^1, \cdots, c^n\}$ where c^i is composed of a group of words $\{x_1^i, ..., x_N^i\}$, we utilize a standard BERT model to encode each utterance c^i as a sentence embedding $\mathbf{u}_c^i$. And then, we apply a BiLSTM on these sentence embedding to obtain the context sequential representation:

$$\mathbf{H}_c^i = \text{BERT}_{base}\{[x_1^i, ..., x_N^i]\},$$
$$\mathbf{u}_c^i = \text{averpool}(\mathbf{H}_c^i),$$
$$\mathbf{h}_c^i = \text{BiLSTM}(\mathbf{u}_c^i, \mathbf{h}_c^{i-1}).$$

Knowledge Transition. We model the knowledge tag transition process with the assistance of Conditional Random Field (CRF) mechanism (Lafferty et al., 2001). We combine a BiLSTM network and a CRF network to form a BiLSTM-CRF model, as shown in Figure 4. This network can efficiently use past input features via a BiLSTM layer and sentence level tag information via a CRF layer. For each BiLSTM cell, it will output the score of each tag. Given a context representation $\mathbf{h}_c^i$, the corresponding tag scores is:

$$score_{i+1}[t^{i+1}] = \text{softmax}(\mathbf{W}_1\mathbf{h}_c^i + \mathbf{b}_1),$$

where $\mathbf{W}_1$ and $\mathbf{b}_1$ are the training parameters. $score_{i+1}[t^{i+1}]$ means the output score of knowledge tag t^{i+1} at the $(i+1)$-th step. CRF layer is capable to model the sequential tag relationship by maximizing a global score $C(t^1, t^2, ...t^n, \theta)$. This global score is the concatenation of a transition score $T[i, j]$ and a matrix of score. $T[i, j]$ is to model the transition probability from i-th tag to j-th for a pair of consecutive steps. The matrix of score is used to record tag transition path along with the context sentences.

$$C(t^1, t^2, ...t^n, \theta) = \sum_{i=1}^{n} T[t^i, t^{i+1}] + \sum_{i=1}^{n} score_{i+1}[t^{i+1}]$$

Therefore, our final selected knowledge tag s_t should be:

$$s_t = \text{argmax}(C(t^1, t^2, ...t^n, \theta)).$$

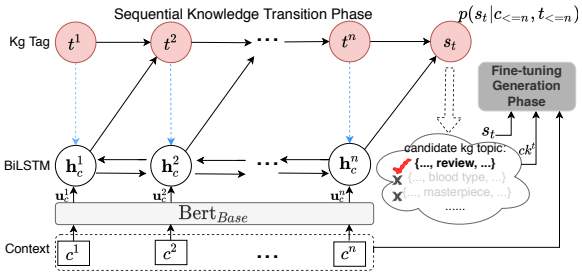

Figure 4: Sequential knowledge transition phase.

Once we get the knowledge tag s_t, we are able to pick out the corresponding knowledge content ck_t from the candidate knowledge set CK. If there are multiple knowledge contents with the same tag s_t, we will apply a coarse-to-fine knowledge matching module to select out the knowledge content with maximum score as ck_t.

Coarse-to-fine Knowledge Matching. To select out the final knowledge content from multiple candidates with the same knowledge tag, we adopt BM25 (Robertson and Zaragoza, 2009), as the supporting coarse-to-fine matching model. Given a knowledge content and dialogue context pair (ck^i, c), the matching model will output a matching score. We will choose the knowledge content with the highest score as the final knowledge content.

Knowledge Transition Loss. In the training phase, we adopt two level knowledge loss to optimize the sequential selection process. Knowledge tag loss $\mathcal{L}_{tag}^{kg}(\theta)$ is a log-likelihood loss to minimized the difference between true tag label and prediction tag label. Knowledge content loss $\mathcal{L}_{cont}^{kg}(\theta)$ is a cross-entropy loss to minimize the divergence between true knowledge sentence and prediction one. Therefore, the total knowledge transition loss is defined to be:

$$\mathcal{L}_{trans}(\theta) = \mathcal{L}_{tag}^{kg}(\theta) + \mathcal{L}_{cont}^{kg}(\theta).$$

3.4 Fine-tuning and Response Generation

The flexible self-attention mask mechanism enables our pre-trained generator to consider the dialogue history in the response generation phase. Given the generated knowledge tag s_t and its corresponding knowledge content ck_t, and the dialogue contexts$\{c^1, \cdots, c^n\}$, the fine-tuning procedure can be carried out by the following training optimization to generate response $y = \{y_1, \cdots, y_N\}$, defined as:

$$\mathcal{L}_{NLL}(\theta) = -\sum_{t=1}^{N} \log p(y_t|y_{<t}, ck_t, s_t, c^1, \cdots, c^n; \theta),$$

The process is shown in the right of the Figure 2. After fine-tuning phase, response can be generated by given selected knowledge tag, corresponding knowledge content, and history dialogue context.

4 Experiments

4.1 Experimental Settings

Dataset. We employ two public knowledge-grounded dialogue benchmarks in our experiments. The structured **DuConv** dataset consists of 29,000 context-response pairs. The corresponding knowledge pool contains 32 different knowledge tags. We randomly divided the corpus into the training, validation and testing set, containing 25,000, 2,000, and 2,000 pairs respectively. The Wizard of Wikipedia (**WoW**) dataset is conducted with 201,999 dialogues about diverse topics. We randomly split this corpus as 18,430 dialogues for training, 1,948 dialogues for validation and 1933 dialogues for test. The test set is split into two subsets: test seen and test unseen. Test Seen contains 965 dialogues on the topics overlapped with the training set, while test unseen contains 968 dialogues on the topics never seen before in training and validation set.

Baselines. We compare our SKT-KG model with several state-of-the-art models, including (i) **Transformer**: a fully self-attention mechanism model (Vaswani et al., 2017), (ii) **MemNet**: The E2E Transformer with memory mechanism (Dinan et al., 2018), which uses a Transformer memory network for knowledge selection and a Transformer decoder for utterance prediction. (iii) **PostKS**: Posterior Knowledge Selection (Lian et al., 2019), which uses the posterior knowledge distribution as a pseudo-label for knowledge selection. (iv) **SLKS**: sequential latent knowledge selection model (Kim et al., 2020), which keeps track of prior and posterior distribution over knowledge and sequentially updated considering contexts in previous turns. we also employ some degraded SKT-KG models to investigate the effect of our proposed pre-trained knowledge-aware response generator mechanisms: **SKT** is the model without pre-trained knowledge-aware response generator, only using the knowledge transition to select the knowledge and then generate the response with transformer decoder.

Parameters Setting. For WoW, we set the vocabulary size to 30,522, as the default setting in BERT [1].

[1]https://github.com/google-research/bert

Dataset	DuConv					
Model	BLEU-1 / 2	Dist-1	Dist-2	Avg.	Ext.	Gre.
Transformer	21.39/11.42	4.67	10.36	51.79	32.07	40.62
MemNet(soft)	22.48/19.95	5.26	13.66	57.28	35.06	41.89
PostKS(fusion)	29.76/21.84	5.84	15.52	55.57	39.54	43.72
SLKS	33.93/24.72	8.40	20.06	59.69	40.31	44.79
SKT	35.79/22.36	9.01	21.47	62.57	46.83	50.11
SKT-KG	**37.80/26.31**	**10.57**	**23.20**	**65.37**	**49.63**	**57.46**

Table 1: Automatic evaluation results on DuConv. The metrics Distinct, Average, Extrema, and Greedy are abbreviated as Dist, Avg., Ext., and Gre., respectively. The best results are highlighted with **bold**.

For DuConv, we set the vocabulary size to 21,128 [2]. To fairly compare our model with all baselines, the number of hidden nodes is all set to 512 and the batch size set to 128. The max length of sentence is set to 30 and the max number of dialogue turns is set to 8. The topic size of LDA for WoW dataset is set as 50. We use Adam (Kingma and Ba, 2014) for gradient optimization in our experiments. The learning rate is set to 0.001. We run all models on the Tesla P40 GPU.

Evaluation Measures. We use both quantitative evaluation and human judgements in our experiments. Specifically, we use the indicators including BLEU-1/2 and distinct-1/2, Embedding metrics (average, extrema and greedy)[3]. We also measure the knowledge selection precision and F1 score between the prediction and ground-truth knowledge. For human evaluation, we randomly sampled 300 generated response and invited six annotators (all CS majored students) to give their rating score based on the relevant, informative and natural of the generated response with respect to the contexts. The rating ranges from 0 to 3 for relevance, informativeness and natural, respectively.

4.2 Experimental Results

4.2.1 Metric-based Evaluation

The metric-based evaluation results are shown in Table 1 and Table 2. From the results, we can see that the sequential knowledge models, i.e., SLKS and our SKT models, perform better than the traditional knowledge-grounded dialogue models, i.e., MemNet and PostKS models, in terms of BLEU and Distinct measures. That's because the sequential characteristic in knowledge is significant and beneficial for the knowledge selection process. Our proposed SKT-KG model obtains good results. Tak-

[2]https://github.com/ymcui/Chinese-BERT-wwm

[3]https://github.com/Maluuba/nlg-eval

Dataset	WoW Test Seen						WoW Test Unseen					
Model	BLEU-1 / 2	Dist-1	Dist-2	Avg.	Ext.	Gre.	BLEU-1 / 2	Dist-1	Dist-2	Avg.	Ext.	Gre.
Transformer	15.76/6.45	2.97	10.72	46.21	34.45	40.98	15.16/5.45	2.45	6.62	41.13	34.71	37.29
MemNet(soft)	16.67/6.67	3.65	11.28	48.23	40.37	44.19	14.72/4.81	2.15	16.18	42.38	35.74	38.53
PostKS(fusion)	17.21/6.98	5.67	21.85	53.36	39.25	45.17	15.61/5.38	2.87	15.18	44.24	38.69	40.38
SLKS	18.91/**7.64**	7.35	26.59	53.98	43.57	51.20	15.91/6.14	2.35	16.59	42.02	39.15	43.66
SKT	19.16/7.32	7.65	27.46	55.99	44.74	47.03	13.50/6.96	3.08	16.04	46.29	39.70	42.43
SKT-KG	**20.62**/7.36	**7.79**	**28.32**	**59.71**	**48.87**	**54.26**	**16.26/6.99**	**3.69**	**16.83**	**52.85**	**41.07**	**45.39**

Table 2: Automatic evaluation results on Wizard of Wikipedia datasets. The metrics Distinct, Average, Extrema, and Greedy are abbreviated as Dist, Avg., Ext., and Gre., respectively. The best results are highlighted with **bold**.

Dataset	Model	F1	klg Acc.
DuConv	MemNet (soft)	15.49	0.22
	PostKS (fusion)	16.38	0.23
	SLKS	17.62	0.26
	SKT	17.75	**0.29**
	SKT-KG	**19.26**	**0.29**
WoW Test Seen	MemNet (soft)	17.23	0.19
	PostKS (fusion)	16.36	0.21
	SLKS	18.91	0.23
	SKT	19.25	**0.26**
	SKT-KG	**19.73**	**0.26**

Table 3: The unigram F1 score and knowledge selection accuracy between SKT-KG and other base-lines on two datasets. The klg stands for knowledge here.

Dataset	Model	Rel	Info	Nat	kappa
DuConv	MemNet(soft)	1.7	1.9	1.6	0.49
	PostKS(fusion)	2.1	2.0	1.7	0.59
	SLKS	1.9	2.2	2.1	0.42
	SKT	2.2	2.1	1.9	0.47
	SKT-KG	**2.3**	**2.6**	**2.3**	0.58
WoW Test Seen	MemNet(soft)	1.6	1.8	1.4	0.45
	PostKS(fusion)	1.7	2.0	1.6	0.51
	SLKS	1.9	**2.3**	1.7	0.49
	SKT	**2.0**	1.9	1.6	0.44
	SKT-KG	**2.0**	2.2	**1.9**	0.46

Table 4: Human evaluation between SKT-KG and other baselines on DuConv and WoW test seen datasets.

ing the BLEU-2 value on the DuConv dataset as an example, the BLEU-2 value of SKT-KG is 26.31, which is better than that of baseline models. The *distinct-2* value of our model is also higher than other baseline models, indicating that our model can generate more diverse responses. For the unigram F1 score of the knowledge selection in Table 3, the F1 score of SKT-KG is 19.26, which is better than other models, showing that our model can extract more relevant and natural knowledge than baseline models. Compared with the ablation model SKT, we find that the pre-trained knowledge-aware response generator in our model can improve *distinct* measure and unigram F1 score, indicating that the model with pre-trained generator has ability to generate more diverse response. We also conducted a significant test. The experimental results show that the improvement of our model is significant in both datasets, i.e., p-value < 0.01. In summary, our SKT-KG model is able to generate higher relevant and more diverse responses than the baselines.

4.2.2 Human Evaluation

The results of human evaluation are shown in Table 4. The rating scores are given to evaluate the relevance, informativeness and natural of the generated responses. From the experimental results, the relevance (Rel), information (Info) and natural (Nat) score for our model is greater than that of MemNet, PostKS and SLKS, indicating that our SKT-KG model is better than the baseline methods. Taking DuConv as an example, the score of relevance and informativeness in SKT-KG are 2.3 and 2.6, respectively, while the SLKS are 2.2 and 2.1, indicating that our model can generate more informative response than SLKS. In addition, for the natural comparison, the score of SKT-KG is 2.3, which is larger than SLKS i.e.,2.1, showing that the high-level knowledge transition is effective for the knowledge-grounded dialogue generation task and our SKT-KG model can generate more natural response with more information. The Kappa (Fleiss, 1971) value demonstrates the consistency of different annotators. We also conducted a significant test, and the improvement of our model is significant on both datasets, i.e., p-value < 0.01.

4.2.3 Case study

To facilitate a better understanding of our model, we present some examples in Figure 5. From the multi-turn dialogues, we can see that the knowledge topic is from ‘ reviews of Mengyao Xi’, to the ‘master work of her’, and then to the ‘ master work of Sui He’. The knowledge tag of ground-truth

Model	DuConv								
	$P@1$	$R@1$	$F@1$	$P@2$	$R@2$	$F@2$	$P@5$	$R@5$	$F@5$
PostKS(fusion)	0.23	0.19	0.21	0.23	0.33	0.27	0.22	0.71	0.34
SLKS	0.26	0.21	0.23	0.25	0.35	0.29	0.25	0.74	0.37
SKT-KG	0.29	0.22	0.25	0.27	0.38	0.32	0.26	0.77	0.39
Model	WoW Test Seen								
	$P@1$	$R@1$	$F@1$	$P@2$	$R@2$	$F@2$	$P@5$	$R@5$	$F@5$
PostKS(fusion)	0.21	0.17	0.19	0.22	0.29	0.25	0.19	0.67	0.30
SLKS	0.23	0.19	0.20	0.21	0.34	0.26	0.19	0.69	0.30
SKT-KG	0.26	0.21	0.23	0.25	0.35	0.29	0.21	0.73	0.33

Table 5: The ranking evaluation of knowledge selection on DuConv and WoW datasets.

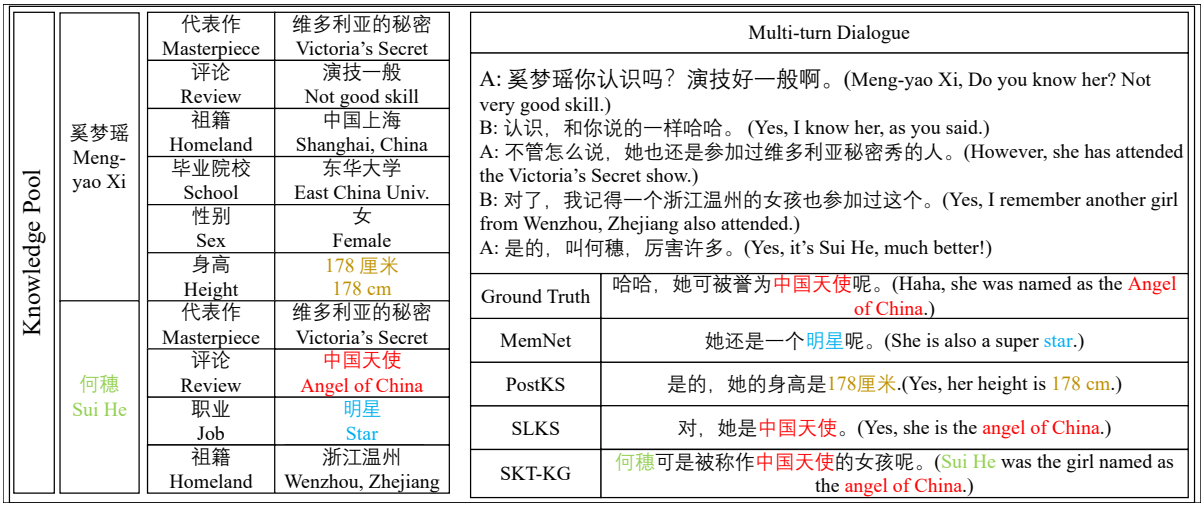

Figure 5: The case of generated response from different models on DuConv.

is the ' reviews of Sui He '. From the generation results, we can see that the sequential-based model performs better than the selection model, i.e., MemNet and PostKS. Taking an example in Figure 5, an un-natural response is generated by MemNet and PostKS, such as 'Area of Sui He ' and ' Height of Sui He'. However, the sequential model can generate more natural and relevant responses, such as ' Yes, she is the angel of China' and 'He Sui was the girl named as the angel of China '. This is mainly because the sequential model is able to locate the ' reviews' knowledge which is more natural for the contexts. Moreover, our high-level transition model with pre-trained knowledge-aware response generator can generate more informative response than SLKS, as shown in Figure 5.

4.3 Analysis on Knowledge Selection

To verify whether the performance improvements are owing to the knowledge transition module, we conduct a further data analysis. Specifically, we randomly sample 300 examples from the DuConv dataset and WoW dataset, to evaluate the performance of the knowledge selection process in baselines and our model. As knowledge-grounded dialogue models will select the relevant knowledge from the candidate knowledge set based on the dialogue contexts, we can treat it as a ranking task. Ranking evaluation measures, such as the precision, recall and F1 score, are used for quantitative evaluations. Then we calculate the precision, recall and F1 score of the top 1,2,5 for PostKS, SLKS and our SKT-KG model. The results are shown in Table 5. We can see that the the sequential knowledge selection models, such as SLKS and SKT-KG, perform better than traditional selection model, i.e., PostKS, validating the effectiveness of sequential knowledge model. These results indicate that our proposed knowledge sequential transition module is capable to select out more relevant knowledge content than baseline models.

5 Conclusion and Future Work

In this paper, we propose a sequential knowledge transition model with knowledge-aware response generator to model the high-level knowledge transition and fully utilize the low-resource knowledge data. SKT-KG models can abstract knowledge into

tags which leads our model easily to apply into both the structured and unstructured knowledge-grounded conversations. Besides, we propose a pre-trained knowledge-aware response generator, aiming at generating a natural sentence based on a given knowledge, to make full use of the limited data. Experimental results on both structured and unstructured knowledge-grounded dialogue datasets show that our SKT-KG model outperforms baseline models. As for future work, we intend to apply variational autoencoder to unstructured dataset, in order to empower models to learn the knowledge topic by themselves.

Acknowledgements

The authors would like to thank all the anonymous reviewers for their constructive comments and suggestions. This work was partially supported by the National Key R&D Program of China under Grants No. 2019AAA0105200, 2016QY02D0405, the Beijing Academy of Artificial Intelligence (BAAI) (No. BAAI2020ZJ0303), the National Natural Science Foundation of China (NSFC) (No. 61722211, 61773362, 61872338, 61906180).

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174.

Hongshen Chen, Zhaochun Ren, Jiliang Tang, Yihong Eric Zhao, and Dawei Yin. 2018. Hierarchical variational memory network for dialogue generation. In *Proceedings of the 2018 World Wide Web Conference*, pages 1653–1662.

Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. The JDDC corpus: A large-scale multi-turn Chinese dialogue dataset for E-commerce customer service. In *LREC*, pages 459–466.

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

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *American Psychological Association*.

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

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. *arXiv preprint arXiv:2002.07510*.

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

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Linxiao Li, Can Xu, Wei Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. Zero-resource knowledge-grounded dialogue generation. *Advances in Neural Information Processing Systems*, 33.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. *arXiv preprint arXiv:1902.04911*.

Ruixue Liu, Meng Chen, Hang Liu, Lei Shen, Yang Song, and Xiaodong He. 2020. Enhancing multi-turn dialogue modeling with intent information for e-commerce customer service. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 65–77. Springer.

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.

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

J Mayo, O White, and Hans J Eysenck. 1978. An empirical study of the relation between astrological factors and personality. *The Journal of Social Psychology*, 105(2):229–236.

Laura Miller. 2014. The divination arts in girl culture.

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

Haoruo Peng, Qiang Ning, and Dan Roth. 2019. Knowsemlm: A knowledge infused semantic language model. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 550–562.

Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2020. Thinking globally, acting locally: Distantly supervised global-to-local knowledge selection for background based conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8697–8704.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

Lei Shen and Yang Feng. 2020. CDL: Curriculum dual learning for emotion-controllable response generation. In *ACL*, pages 556–566.

Lei Shen, Yang Feng, and Haolan Zhan. 2019. Modeling semantic relationship in multi-turn conversations with hierarchical latent variables. In *ACL*, pages 5497–5502.

Lei Shen, Haolan Zhan, Xin Shen, and Yang Feng. 2021. Learning to select context in a hierarchical and global perspective for open-domain dialogue generation. *arXiv preprint arXiv:2102.09282*.

Hui Su, Xiaoyu Shen, Sanqiang Zhao, Zhou Xiao, Pengwei Hu, Randy Zhong, Cheng Niu, and Jie Zhou. 2020. Diversifying dialogue generation with non-conversational text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7087–7097.

Yajing Sun, Yue Hu, Luxi Xing, Jing Yu, and Yuqiang Xie. 2020. History-adaption knowledge incorporation mechanism for multi-turn dialogue system. In *AAAI*, pages 8944–8951.

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 *Advances in neural information processing systems*, pages 5998–6008.

Pavlos Vougiouklis, Jonathon Hare, and Elena Simperl. 2016. A neural network approach for knowledge-driven response generation. In *COLING*.

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.

Jun Xu, Haifeng Wang, Zhengyu Niu, Hua Wu, and Wanxiang Che. 2020. Knowledge graph grounded goal planning for open-domain conversation generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9338–9345.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Hainan Zhang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. 2018a. Reinforcing coherence for sequence to sequence model in dialogue generation. In *IJCAI*, pages 4567–4573.

Hainan Zhang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. 2018b. Reinforcing coherence for sequence to sequence model in dialogue generation. In *International Joint Conference on Artificial Intelligence*, pages 4567–4573.

Hainan Zhang, Yanyan Lan, Liang Pang, Hongshen Chen, Zhuoye Ding, and Dawei Yin. 2020. Modeling topical relevance for multi-turn dialogue generation. *arXiv preprint arXiv:2009.12735*.

Hainan Zhang, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. 2019. Recosa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3721–3730.

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020. Low-resource knowledge-grounded dialogue generation. *arXiv preprint arXiv:2002.10348*.

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.