TAKE: Topic-shift Aware Knowledge sElection for Dialogue Generation

Chenxu Yang^{1,2}, Zheng Lin^{1,2}*, Jiangnan Li^{1,2}, Fandong Meng³, Weiping Wang¹, Lanrui Wang^{1,2}, Jie Zhou³

¹Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China ²School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

³Pattern Recognition Center, WeChat AI, Tencent Inc, China

{lijiangnan, linzheng, wangweiping}@iie.ac.cn

{yangchenxu22,wanglanrui22}@mails.ucas.ac.cn

{fandongmeng,withtomzhou}@tencent.com

Abstract

Knowledge-grounded dialogue generation consists of two subtasks: knowledge selection and response generation. The knowledge selector generally constructs a query based on the dialogue context and selects the most appropriate knowledge to help response generation. Recent work finds that realizing who (the user or the agent) holds the initiative and utilizing the role-initiative information to instruct the query construction can help select knowledge. It depends on whether the knowledge connection between two adjacent rounds is smooth to assign the role. However, whereby the user takes the initiative only when there is a strong semantic transition between two rounds, probably leading to initiative misjudgment. Therefore, it is necessary to seek a more sensitive reason beyond the initiative role to help refine the history information used to construct the query. To address the above problem, we propose a Topic-shift Aware Knowledge sElector(TAKE). Specifically, we first annotate the topic shift and topic inheritance labels in multiround dialogues via distant supervision. Then, we alleviate the noise problem in pseudo labels through curriculum learning and knowledge distillation. Extensive experiments on WoW show that TAKE performs better than strong baselines.¹

1 Introduction

Due to the dull response generation problem in the general open-domain dialogue generation technology, an increasing number of researchers focus on knowledge-grounded dialogue generation (KGDG) (Ghazvininejad et al., 2017; Li et al., 2019; Chen et al., 2020a; Zhan et al., 2021a). By connecting the external knowledge base with the generation model as supplement information, the generated response becomes more engaging and informative.

	Topic: French Bulldog
User	(1) I love my French bulldog!
	(2) Aww, I bet your dog is so cute.
Agent	The French Bulldog is a small breed
	and also known as the Frenchie.
	<french bulldog's="" nickname=""></french>
Usor	(3) Yes, it is. What <u>color</u> is a bulldog
User	usually?
	(4) They are a cross between bulldogs
Agent	and ratters. Most are white or black.
	<french bulldog's="" color=""></french>
User	(5) I see, do you own any pets?
	(6) I have a pet snake. His name
Agent	is Slinky.
	<snake></snake>
User	(7) Cool! Is it safe to have snake pets?
	(8) Yes, if you have the right enclosure.
	I have one that likes to eat prey much
Agent	larger than his head. I feed him rats.
	<snake></snake>

Table 1: An example of topic shift in WoW dataset. The utterances (2) to (4) exhibit a sub-topic shift: from dog's nickname to dog's color; the utterances (4) to (6) exhibit an obvious topic shift: from dog to snake; the utterances (6) to (8) exhibit a topic inheritance.

Knowledge selection plays a vital role in KGDG task (Meng et al., 2021). Since one can choose any reasonable knowledge to carry on the conversation, one-to-many relations exist between dialogue context and knowledge (Kim et al., 2020). Thus, selecting the most appropriate knowledge in the vast knowledge pool becomes a significant challenge. And most of the existing methods pay close attention to the design of the knowledge selector.

Some methods attempt to improve the accuracy of knowledge selection by discovering more features in the dialogue context (Meng et al., 2020; Zheng et al., 2020) or introducing extra posterior knowledge (Lian et al., 2019; Kim et al., 2020; Chen et al., 2021). These methods often directly

^{*} Zheng Lin is the corresponding author.

¹The code is available at https://github.com/ iie-ycx/COLING2022-TAKE.

take the whole dialogue context as the input of the knowledge selector, ignoring that different parts of the context play different roles. Considering this, Meng et al. (2021) propose to decouple the knowledge selector according to the different part of the input and only keep part of the history information by introducing mixed-initiative (userinitiative and agent-initiative) characteristics. In their method, the user takes the initiative when the knowledge connection between two adjacent rounds is unsmooth. Such a judgement only works when there is a strong semantic transition between two rounds. However, the conversational direction can be changed by the user when the user shifts the topic from one to another relevant one. For example, the utterance flow $(2) \rightarrow (3) \rightarrow (4)$ in Table 1 shows that the topic shifts from *dog's nickname* to dog's color by the user, which means that the user dominates the dialogue direction. Nevertheless, the knowledge connection here is smooth because the knowledge is still relevant to dogs, leading to an agent-initiative judgement. The misjudgement will make the model choose the improper part of (agentrelated) history information to select knowledge.

To address the above problem, we bring the topic into multi-turn knowledge selection. Through our observation, we find that the topic shift and topic inheritance affect knowledge selection deeply. For the topic shift, it is generally caused by the active user's frequent questioning and the model should select knowledge according to the current user utterance. For the topic inheritance, it is mainly caused by relatively passive users who agree with what the agent says, and the model needs to find some relative topics to continue the conversation according to previously selected knowledge. We obtain the topic shift label via distant supervision (Mintz et al., 2009), where we regard the retrieving entity as the topic word. Considering that there may exist noises in the pseudo labels, we further alleviate their negative effects through curriculum learning and knowledge distillation (see section 2.6 for details).

Our contributions in this paper are as follows:

- For the KGDG task, we find that the topic shift triggers knowledge alteration, and propose a Topic-shift Aware Knowledge sElector (TAKE) to better locate the relevant parts from the dialogue history at an opportune moment.
- To overcome the noisy label problem introduced by distant supervision, we optimize the

topic-shift aware knowledge selector through curriculum learning and knowledge distillation, which can effectively alleviate the negative influence of pseudo topic labels.

• Experimental results on WoW dataset show that compared with strong baselines, TAKE not only selects knowledge more accurately especially on the unseen test set, but also generates more informative responses on both automatic and human evaluation metrics.

2 Approach

2.1 Task Formulation

Suppose we have a *t*-rounds conversation $C = \{(X_t, Y_t)\}, t = 1, 2, ..., |C|$, where X_t and Y_t are the utterances of the user and the agent at turn *t* respectively. In each turn, before the dialogue agent generates responses, the model is externally connected with a knowledge pool $\mathcal{K}_t = \{K_1^t, K_2^t, \ldots, K_D^t\}$ which contains *D* pieces of knowledge. Given the current user utterance X_t , the dialogue history $\{X_i, Y_i\}_{i=1}^{t-1}$, the previous golden knowledge $\{K_i'\}_{i=1}^{t-1}$ and the current knowledge pool \mathcal{K}_t , our goal is to select the most appropriate knowledge K_t' from the pool and make use of the selected knowledge K_t^s to generate the response $Y_t = (y_1^t, y_2^t, \ldots, y_{|Y_t|}^t)$.

2.2 Overview of TAKE

As shown in Figure 1, our model TAKE contains three components: Mixed Encoder, Topic-aware Knowledge Selector and Decoder. In the following subsections, we first introduce the three components in section 2.3, 2.4 and 2.5. Next, we present how to utilize curriculum learning and knowledge distillation to alleviate the problem of noisy pseudo labels obtained by distant supervision in section 2.6. Finally, we detail our training strategy and loss functions.

2.3 Mixed Encoder

We take BERT as the backbone of Encoder. At turn t, given the user utterance X_t and the knowledge pool $\mathcal{K}_t = \{K_1^t, K_2^t, \dots, K_D^t\}$, we mix the user utterance and the knowledge sentences following (Zhao et al., 2020b). Specifically, we concatenate X_t with [CLS] [SEP] token in BERT and the candidate knowledge K_i^t to acquire M_i^t . In this way, we can better use the multi-layer bidirectional attention mechanism in BERT to allow the dialogue context interact sufficiently with knowledge candidates.



Figure 1: Architecture of the proposed model.

$$M_i^t = [\text{CLS}]X_t[\text{SEP}]K_i^t \tag{1}$$

We then take D concatenated context-knowledge pairs $M_t = \{M_1^t, M_2^t, \dots, M_D^t\}$ as the input of the Encoder. Tokens in M_t will be encoded as word representations $[\mathbf{H}^{\text{CLS}_i^t}; \mathbf{H}^{X_i^t}; \mathbf{H}^{\text{SEP}_i^t}; \mathbf{H}^{K_i^t}]$. After that, we obtain sentence representations via mean pooling (Cer et al., 2018):

$$\mathbf{h}^{X_i^t} = \text{pooling}(\mathbf{H}^{X_i^t}),$$

$$\mathbf{h}^{K_i^t} = \text{pooling}(\mathbf{H}^{K_i^t})$$
(2)

In order to get a unified sentence representation of the user utterance, we integrate the D representations through additive attention mechanism and obtain \mathbf{h}^{X_t} :

$$s_i = v_u^{\top} \tanh(W_u \mathbf{h}^{X_i^t})$$

$$\alpha^t = \mathrm{SM}(\{s_i\}_{i=1}^D), \mathbf{h}^{X_t} = \sum_{i=1}^D \alpha_i^t \mathbf{h}^{X_i^t}$$
(3)

where v_u and W_u are trainable weights, and SM means softmax function.

2.4 Topic-aware Knowledge Selector

The knowledge selector module we designed is composed of three parts: Topic Shift Discriminator, Topic-shifted Knowledge Selector and Topicinherited Knowledge Selector. We name the latter two networks as Sub-KS for simplicity.

The topic shift discriminator can judge whether the topic shift or the topic inheritance will occur at the current turn and then choose one selector in Sub-KS. According to the previous analysis, in the case of topic shift, it is more likely that the user mentions a new topic. Hence the topic-shifted knowledge selector makes full use of the current user utterance to construct query vectors. Otherwise, the topicshifted knowledge selector inherits a topic from the previous conversation and selects a knowledge under the topic.

Due to BERT's NSP pre-training scheme (Devlin et al., 2018), the [CLS] token is endowed with the ability to extract semantic information in sentences. We use a feedforward layer to extract the semantic associative information between the current user utterance and the knowledge candidates as:

$$\mathbf{u}_{i}^{t} = \operatorname{Relu}(\operatorname{FC}(\mathbf{H}^{\operatorname{CLS}_{i}^{\iota}})) \tag{4}$$

where FC means fully connected layers.

Besides, we use the attention mechanism on knowledge candidates similar to function (3) and then extract the differential features between them through a multilayer perceptron(MLP).

$$\mathbf{h}^{K_t} = \operatorname{attention}(\{\mathbf{h}^{K_i^t}\}_{i=1}^D),$$

$$\mathbf{e}_t^k = \operatorname{Relu}(\operatorname{FC}(\mathbf{h}^{K_t}))$$
(5)

Topic-shifted Knowledge Selector: Given the current user utterance representation \mathbf{h}^{X_t} , the candidate knowledge representations $[\mathbf{h}^{K_1^t}, \dots, \mathbf{h}^{K_D^t}]$, and the representations mentioned above \mathbf{u}_i^t , \mathbf{e}_t^k , we construct query vector and key vector as:

$$\mathbf{Q}_{sh} = \mathrm{MLP}([\mathbf{h}^{X_t}; \mathbf{e}_t^k])$$

$$\mathbf{K}_i^{sh} = \mathrm{MLP}([\mathbf{u}_i^t; \mathbf{h}^{K_i^t}])$$

$$\mathbf{K}_{sh} = [\mathbf{K}_1^{sh}; \dots; \mathbf{K}_D^{sh}]$$
(6)

Given \mathbf{Q}_{sh} and \mathbf{K}_{sh} , the topic-shifted knowledge selector predicts the distributions over the knowledge pool \mathcal{K}_t by additive attention:

$$P(\mathcal{K}_t|S) = \mathrm{SM}(v_0^{\top} \mathrm{tanh}(W_0 \mathbf{K}_{sh} + U_0 \mathbf{Q}_{sh}))$$
⁽⁷⁾

where v_0, W_0 and U_0 are trainable parameters.

Topic-inherited Knowledge Selector: Following (Meng et al., 2021), we apply a stack of transformer encoder blocks with positional embeddings to integrate the previously selected knowledge and extract the inherited topic $\mathbf{h}_{trans}^{K'_{t-1}}$:

$$[\mathbf{h}_{trans}^{K_{1}'};\ldots;\mathbf{h}_{trans}^{K_{t-1}'}] =$$
TransformerEncoder([$\mathbf{h}^{K_{1}'};\ldots;\mathbf{h}^{K_{t-1}'}]$)
(8)

Then, given the representation of inherited topic $\mathbf{h}_{traps}^{K_{t-1}}$, the candidate knowledge representations $[\mathbf{h}^{K_1^t}, \ldots, \mathbf{h}^{K_D^t}]$ and the representation of knowledge difference \mathbf{e}_t^k , we construct query vector and key vector as:

$$\mathbf{Q}_{inh} = \mathsf{MLP}([\mathbf{h}_{trans}^{K_{t-1}}; \mathbf{e}_{t}^{k}])$$

$$\mathbf{K}_{i}^{inh} = \mathsf{MLP}(\mathbf{h}_{i}^{K_{i}^{t}})$$

$$\mathbf{K}_{inh} = [\mathbf{K}_{1}^{inh}; \dots; \mathbf{K}_{D}^{inh}]$$
(9)

Similarly, the topic-inherited knowledge selector predicts the distributions $P(\mathcal{K}_t | inherit)$ by:

$$P(\mathcal{K}_t|I) = \mathbf{SM}(v_1^{\top} \tanh(W_1 \mathbf{K}_{inh} + U_1 \mathbf{Q}_{inh}))$$
(10)

Topic Shift Discriminator: There are two topic shift discriminators at the training stage: teacher topic shift discriminator and student topic shift discriminator. The former is provided with the current golden knowledge as posterior information, and it generates soft labels of topic shift, which can guide the student model to distinguish noises. In section 3.3 we will explain how it works in detail. The latter generates hard labels indicating which Sub-KS works, and this corresponds to the 0/1 switch in Figure 1.

Given the integration of the previously selected knowledge $\mathbf{h}_{trans}^{K_{t-1}'}$ and the current-turn user utterance representation \mathbf{h}^{X_t} , we first extract the topic information in the current user utterance \mathbf{e}_t^x by a multilayer perceptron(MLP). Then we construct two discriminators based on whether they have access to posterior information as follows:

Teacher module:

$$\mathbf{v}_{\mathcal{T}} = [\mathbf{h}^{K'_t}; \mathbf{e}^x_t; \mathbf{h}^{K'_{t-1}}_{trans}; \mathbf{e}^x_t - \mathbf{h}^{K'_{t-1}}_{trans}; \mathbf{e}^x_t \odot \mathbf{h}^{K'_{t-1}}_{trans}]$$
$$P_{\mathcal{T}}(\mathcal{D} = S) = \text{Sigmoid}(\text{FC}(\mathbf{v}_{\mathcal{T}}))$$
(11)

Student module:

$$\mathbf{v}_{\mathcal{T}} = [\mathbf{e}_{t}^{x}; \mathbf{h}_{trans}^{K_{t-1}}; \mathbf{e}_{t}^{x} - \mathbf{h}_{trans}^{K_{t-1}'}; \mathbf{e}_{t}^{x} \odot \mathbf{h}_{trans}^{K_{t-1}'}]$$

$$P_{\mathcal{S}}(\mathcal{D} = S) = \text{Sigmoid}(\text{FC}(\mathbf{v}_{\mathcal{S}}))$$
(12)

where \odot denotes element-wise product, and $\mathcal{D} = S$ means discrimination result is topic shift.

We train both of the two modules with Cross Entropy loss, $\hat{\mathbf{y}}_t$ represents the topic shift labels obtained by distant supervision:

$$L(\mathcal{D}) = -\frac{1}{|C|} \sum_{t=1}^{|C|} \hat{\mathbf{y}}_t \log(P(\mathcal{D})) \qquad (13)$$

2.5 Decoder

We take GPT-2 as the backbone of Decoder. Similar to (Zhao et al., 2020b; Zheng and Huang, 2021), we define new tokens in the dictionary of GPT-2 Tokenizer: "<context>", "<response>" and "<knowledge>". These tokens are treated as segment embeddings to mark different information components.

At the training stage, the inputs of the decoder are the concatenation of the dialogue context, the golden knowledge and the responses. The training loss is the Cross Entropy on the responses. At the inference stage, given the selected knowledge K_{sel}^t and the dialogue context $\{(X_1, Y_1), \ldots, (X_{t-1}, Y_{t-1}), X_t\}$, the decoder synthesizes the two parts to generate the current response until <eos>.

$$L_g = -\frac{1}{|Y_t|} \sum_{i=1}^{|Y_t|} \log(P(y_i^t | X_{\le t}, Y_{< t}, K_t', y_{< i}^t))$$
(14)

2.6 Pseudo Label Learning

Our findings on topic shift in multi-round conversation suggest that topic shift helps knowledge selection. However, we still face a lack of labels. Inspired by distant supervision, we regard the entity words of retrieving knowledge candidates as topic words and obtain topic shift labels \mathcal{D}' . After that, we alleviate the noise problem in pseudo labels through curriculum learning and knowledge distillation. In this subsection, we describe our methods in detail.

Distant Supervision: During the construction of the KGDG dataset, the candidate knowledge is retrieved from Wikipedia by entity words through search engines. The entity word can retrieve the corresponding knowledge, which indicates that it highly summarizes the content of the knowledge sentence. However, the existing methods ignore this useful information. According to (Brown and Yule, 1983), the topic is the most frequently used term. We make a hypothesis according to the idea of distant supervision (Mintz et al., 2009): the entity used to search candidate is the topic word of each knowledge sentence. Then, we mark topic shift labels according to the topic words. If the topic words of the current round appeared in the dialogue context, we mark the current round as topic inheritance with label 0; if the entity words of the current round have not appeared in the dialogue context, we mark the current round as topic shift with label 1. We will prove that the pseudo labels acquired in this way are instructing in section 3.6.

Noisy Label Learning: The hypothesis mentioned above is relatively strong. Therefore, we adopt curriculum learning (Bengio et al., 2009) and knowledge distillation (Hinton et al., 2015) methods to alleviate the noises and further optimize the model. Specifically, with the increase of training steps, our model gradually chooses the Sub-KS based on the output of the discriminator to reduce the dependence on the topic shift label D'.

$$L_{CL} = -\frac{1}{|C|} \sum_{t=1}^{|C|} (p_i \log \left(P\left(\mathcal{K}_t\right) P\left(\mathcal{D}'\right) \right) + (1-p_i) \log \left(P\left(\mathcal{K}_t\right) P\left(\mathcal{D}_S\right) \right))$$
(15)

where p_i decreases with the training steps.

Besides, we distill knowledge between the teacher model and the student model through KL divergence loss:

$$L_{distill} = D_{KL}(P_{\mathcal{T}}||P_{\mathcal{S}}) \tag{16}$$

We train TAKE's knowledge selector in two stages:

Stage I:
$$L_{\mathcal{K}_1} = L_{ks} + \alpha(L_{\mathcal{T}})$$

Stage II: $L_{\mathcal{K}_2} = L_{CL} + \alpha(L_{\text{distill}} + L_{\mathcal{S}})$
(17)

and the original KS loss is defined as:

$$L_{ks} = -\frac{1}{|C|} \sum_{t=1}^{|C|} \hat{\mathbf{y}}_t \log\left(P\left(\mathcal{K}_t\right) \mid P\left(\mathcal{D}'\right)\right) \quad (18)$$

3 Experiments

3.1 Experimental Setup

Dataset. There are dozens of datasets chosen for evaluating the KGDG task before (Moghe et al., 2018; Zhou et al., 2018; Dinan et al., 2018). We take the most challenging one Wizard of Wikipedia(WoW) for experiments. The WoW data is obtained from a crowdsourcing data collection website. In data collection, the user side plays the role of the apprentice, and the agent side plays the role of the wizard. The wizard has access to the knowledge retrieved from Wikipedia as groundsource to generate informative responses, while the apprentice prefers speaking common utterances. In the WoW dataset, there are nearly 67 pieces of knowledge on average in a knowledge pool. The WoW dataset consists of 22,311 dialogues with 201,999 turns divided into training set/validation set/test set. The test set is further divided into test seen set and test unseen set. The conversation topics of the test seen set have appeared in the training set, while the topics of the test unseen set are brand-new. The latter contains out-of-domain data which is more challenging.

Baseline Models. We compare our TAKE model with several SOTA models, including:

MemNet: A model proposed by Dinan et al. (2018), which is regarded as the most basic baseline in the KGDG task.

SKLS: Kim et al. (2020) design the sequential latent knowledge selection model according to the idea of conditional variational auto-encoder.

DukeNet: Meng et al. (2020) design a knowledge selection network that takes knowledge tracking and knowledge transfer as a pair of dual tasks to provide feedback to each other.

DiffKS: A model considered the difference of knowledge between the two adjacent rounds of dialogue, proposed by Zheng et al. (2020). We take the decoupled version as one baseline.

KnowledGPT: Zhao et al. (2020b) design a joint training strategy, which uses the combination of RL and CL to improve the knowledge selection module and generation module. It takes GPT-2 as the decoder.

MIKe: Meng et al. (2021) introduce the concept of mixed-initiative into knowledge-grounded dialogue generation and train the initiative discriminator with a self-supervised learning strategy.

CoLV: A model proposed by Zhan et al. (2021a), the authors propose a collaborative latent variable model to integrate the diversity of KS and DG simultaneously in separate yet collaborative latent spaces.

MIKe + GPT2: We reinforce the MIKe model with GPT-2 as a strong baseline, and the KS module remains unchanged.

Evaluation Metrics. For automatic evaluation, we evaluate KS with accuracy and evaluate response generation quality with sentence-

Model	WoW Test Seen							WoW Test Unseen				
Widder	BLEU-1	BLEU-4	RG-1	RG-L	METEOR	ACC	BLEU-1	BLEU-4	RG-1	RG-L	METEOR	ACC
MemNet	17.2	1.61	24.1	17.0	15.5	23.9	13.7	0.6	21.7	15.6	13.1	14.0
SKLS	18.9	1.8	24.5	17.6	16.0	26.8	17.3	1.1	21.0	16.1	13.7	18.3
DukeNet	18.6	2.6	25.4	18.8	17.3	26.2	16.3	1.8	23.2	16.9	15.4	20.1
DiffKS	18.8	2.2	24.8	17.9	16.8	25.6	17.4	1.7	23.6	16.8	14.7	19.8
MIKe	19.1	2.8	25.9	19.2	18.3	28.2	17.6	2.1	24.2	17.8	16.0	21.5
CoLV	-	2.9	20.6	-	-	30.1	-	2.1	19.7	-	-	18.9
KnowledGPT	19.5	-	24.7	-	-	28.0	17.7	-	22.3	-	-	25.4
MIKe + GPT2	20.4	3.3	26.7	20.2	19.4	28.2	18.8	2.5	25.1	18.6	17.4	21.5
TAKE	20.8	3.6	27.1	20.5	19.9	28.8	20.1	3.3	26.2	19.7	18.9	25.8

Table 2: Automatic Evaluation results on Wizard of Wikipedia.

Model	WoW	Test Seen		WoW Test Unseen			
WIOdel	informativeness	coherence	fluency	informativeness	coherence	fluency	
Dukenet	1.63	1.96	1.69	1.57	1.89	1.61	
MIKe	1.66	1.97	1.62	1.61	1.90	1.86	
MIKe+GPT-2	1.84	2.09	2.05	1.70	2.15	2.27	
TAKE	1.88	2.14	2.05	1.90	2.21	2.29	

Table 3: Human Evaluation results on Wizard of Wikipedia. The improvement of TAKE to the best baseline (MIKe+GPT-2) is statistically significant (t-test with p-value < 0.05).

level BLEU-1 (Papineni et al., 2002), BLEU-4, ROUGE-1 (Lin, 2004), ROUGE-L and METEOR (Denkowski and Lavie, 2014). These metrics have been widely used in generation tasks before. For human evaluation, We randomly sample 50 responses in test seen set and 50 responses in test unseen set. By labeling manually, we find the proportion of the topic shift turn is close to the topic inheritance turn among these samples. Then we invite five knowledgeable annotators to score these samples in $\{0,1,2,3\}$ considering the following three aspects: context coherence, fluency and informativeness(which response contains more knowledge and looks more informative). We compute Fleiss' kappa value (Fleiss, 1971) among different annotators to measure their agreement.

Implementation Details. We use PyTorch (Paszke et al., 2019) framework to implement our model. For the implementation of pre-training models BERT(110M) and GPT-2(117M), we utilize the open-source Hugging Face transformers (Wolf et al., 2020). The whole model is optimized with Adam (Kingma and Ba, 2014) algorithm and gradient clipping with a maximum gradient norm of 0.4. We use the gradient accumulation method (accumulation number is 16), and preprocess the knowledge pool by limiting the number of candidate knowledge to 32 and retaining the golden knowledge at the training stage to save GPU memory. The batch size is 2 for training KS and 4 for training DG. The learning rates are 1e-5 for BERT; 6e-5 for the topic-

aware knowledge selector; 3e-5 for GPT-2. We adopt the linear scheduler with a warm-up strategy for the training of Bert and the knowledge selector. It takes total ten epochs for training stages I and II and five epochs for training the GPT-2 Decoder. The weight α in multi-task learning is set to 0.5. Our model is trained on one NVIDIA Geforce RTX 3090 GPU. For other settings, such as the hidden size, dropout rate, sentence length and so on, we keep consistent with MIKe.

3.2 Experimental Results

Table 2 demonstrates automatic evaluation results on WoW. Our model outperforms the typical baseline methods Memnet, SKLS, Dukenet and Diffks remarkably. These methods treat the dialogue context equally to construct query vectors, and the cursory construction hurts knowledge selection, thus leading to terrible generation results. In terms of KS performance, although TAKE has no obvious advantage in the test seen set compared with the strong baselines MIKe and Colv, it has a notable improvement in the test unseen set (+4.3% MIKe,+6.9% CoLV). We think that the promotion comes from a better location of history information of our method to conduct attention mechanism. Besides, TAKE has a significant improvement over all generation metrics, which indicates that TAKE can generate more informative and engaging responses. For fairness, we transplant the GPT-2 module in TAKE to MIKe to make a further com-

Model			WoW Te	st Seen			V	VoW Tes	t Unseen	l		
Widder	BLEU-1	BLEU-4	RG-1	RG-L	METEOR	ACC	BLEU-1	BLEU-4	RG-1	RG-L	METEOR	ACC
TAKE	20.84	3.58	27.14	20.45	19.88	28.75	20.12	3.25	26.16	19.74	18.85	25.78
-CL	20.65	3.41	27.08	20.24	19.69	28.22	19.95	3.17	26.05	19.62	18.76	25.11
-CL -distill	20.53	3.25	26.86	20.12	19.63	27.83	19.82	3.17	25.96	19.49	18.58	24.75
-label Soft	20.55	3.27	26.91	20.15	19.62	27.92	19.68	2.99	25.90	19.34	18.42	23.37
-label Hard	20.12	3.14	26.34	19.75	19.01	25.43	19.36	2.94	25.35	19.10	18.08	22.32

Table 4: Ablation study on the WoW dataset. -CL denotes removing the curriculum learning method. -distill denotes removing the teacher model and knowledge distillation method. -label Soft denotes removing the pseudo topic shift label for supervising and connecting the discriminator and Sub-KS with probability. -label hard denotes removing the pseudo topic shift label for supervising, and the discriminator still chooses Sub-KS through the gumbel-softmax trick.

parison. The results further emphasize that our topic-aware knowledge selector has stronger ability of KS than the mixed-initiative knowledge selector.

The human evaluation is shown in Table 3. The kappa values are between 0.65 and 0.75, denoting substantial agreement among the annotators .Thanks to GPT-2 model, TAKE is superior to the typical baselines on all metrics. As for the strong baseline MIKe+GPT2, although it is comparable with TAKE on fluency, our model exceeds it on informativeness and context coherence a lot. We think the reason is that TAKE selects knowledge which is more coherent with the dialogue context.

3.3 Ablation Study

In order to clarify the source of performance improvement in TAKE, we conduct ablations by removing particular modules from TAKE. The ablation results are shown in Table 4, which denotes all components are beneficial for TAKE. The two methods of noise alleviation in section 2.6 can improve the performance in the inference stage by making the model adapt to noises introduced by inaccurate discrimination in advance. Besides, because the topic transfer label is binary, the curriculum learning method can neutralize the noises in the pseudo labels. With the increase of the training rounds, TAKE has more possibilities on discriminating noisy samples which avoids the overfitting on wrong labels. The teacher model with posterior information learns part of the noises in the pseudo labels in advance and guides the student model to correct these samples in the form of soft labels. For the last two lines of the experiment, we find that the model tends to deteriorate without instructing labels. Only one of the two sub-KS is activated, and the other is idle. The reason is that the topic shift discriminator is not supervised by labels, hence only one of the two sub-KS has been fully trained. This set of experiments proves the importance of

the pseudo labels.

3.4 Case Study

To better evaluate the performance of response generation, we randomly select some examples from the WoW dataset generated by DukeNet, MIKe, KnowledGPT and TAKE to make comparisons. In Table 5, the user asked when Instagram appeared at the current turn. Dukenet and the knowledGPT model do not capture this vital information to construct queries, so they select the wrong knowledge. MIKe mistakenly judges that the current round is an agent initiative dialogue because the connection between the history and current responses is smooth. Consequently, it selects knowledge about Instagram usage based on the previously selected knowledge. Only TAKE constructs the query vector based on the current user utterance and selects the most appropriate external knowledge. We post the remaining topic-inherited example in appendix A.2.



Figure 2: Recall experiment of TAKE on WoW dataset

3.5 Multi-Knowledge Integration Performance

By conducting experiments on Recall of KS and multi-sentence knowledge integration, we find that the one-to-many relations between the dialogue context and knowledge occur more frequently during topic inheritance. Under the framework we

	Topic: Instagram
K_0	no passages used.
K_1	Users can apply various digital filters to their images and add locations through geotags.
K_2	It was created by Kevin Systrom and Mike Krieger, and launched in October 2010 as a free mobile app exclusively for the iOS operating system.
K_3	They can add hashtags to their posts, linking the photos up to other content on Instagram featuring the same subject or overall topic.
context	(1)Are you familiar with Instagram? (2) Oh yes, I love going on Instagram! (3) I think it makes the images look much more professional. (4) Yeah, they have many neat features. I wonder when it was started?
DukeNet	$(K_0 \times)$ I'm not sure but I know that they are called " x ".
KnowledGPT	$(K_0 \times)$ I'm not sure but I know that it was made in the world.
MIKe	$(K_1 \times)$ I'm not sure, but I know that users can apply digital filters through geotags.
TAKE	$(K_2\checkmark)$ It was launched in October 2010 as a free mobile app exclusively for the iOS system.

Table 5: An example for case study. K_2 is the golden knowledge. Only TAKE focuses on the current user utterance to select the golden knowledge and integrates it in response.

Model	We	W Test	Seen	WoW Test Unseen			
WIGGET	BLEU-1	RG-1	METEOR	BLEU-1	RG-1	METEOR	
MIKe+GPT2	20.44	26.74	19.42	18.77	25.06	17.40	
MIKe+GPT2 R@2	20.28	26.65	19.27	18.76	24.92	17.47	
TAKE	20.84	27.14	19.88	20.12	26.16	18.85	
TAKE R@2	20.65	27.00	20.11	19.92	26.15	18.85	
TAKE Inh R@2	20.89	27.24	20.14	20.21	26.26	18.94	

Table 6: Top-2 knowledge integrated evaluation results on WoW. R@2 denotes integrating two knowledge; TAKE Inh R@2 denotes TAKE's topic-inherited KS integrates two knowledge.

proposed, the topic-inherited selector inherits topics in previous rounds. Even though TAKE does not select the golden knowledge, it chooses rational knowledge with closing topics. The golden knowledge probably ranks at the k-th positions (k>1). As shown in Figure 2, with the increase of Recall, the curve of the topic-inherited selector rises steeper, and the improvement of KS comes from topic-inherited rounds.

Apart from that, we find our topic-aware framework is very suitable for explicit multi-knowledge integration. Studies have shown that increasing the number of knowledge integrated into dialogue generation blindly hurts the generation performance (Bruyn et al., 2020), which can be confirmed by the first four lines in Table 6. This is because there is roughly no relationship between the rank-1 and rank-2 items of knowledge. However, the generation results improve if we recall top-2 knowledge only for topic-inherited rounds. The phenomenon explains that TAKE is robust, and it can utilize multi-knowledge effectively.

3.6 Analysis of Pseudo Label and Noisy Label Learning

To prove the effectiveness of our pseudo labels and noisy label learning methods, we further conduct experiments by replacing the output of the discriminator with random 0/1 labels or different proportion of pseudo labels at the inference stage. Table 7 and Figure 3 exhibit the results. If we use random topicshift label, the Sub-KS performs terribly with the wrong part of history information. By comparison, the last line in Table 7 denotes that if TAKE can learn to discriminate topic shift the same as pseudo labels, the performance will improve much better. Apart from that, the rising trends in Figure 3 apparently shows that the more proportion of pseudo labels the model obtains, the better results it performs. In our experiments, TAKE's topic shift discriminator can predict about 78 percent pseudo labels after training. However, its knowledge selection accuracy is significantly higher than 78 percent pseudo-labels-given experiment during inference on both the seen and unseen test set. The phenomenon demonstrates that the curriculum learning and knowledge distillation methods alleviate the noisy pseudo label problem, and then get fine performance which can only be reached under a higher proportion of pseudo labels.

4 Related Work

Knowledge-grounded dialogue generation In recent years, the KGDG task has been a hot spot of research, and many new datasets have emerged (Zhou et al., 2018; Dinan et al., 2018; Eric et al., 2021; Komeili et al., 2022). The existing work of

Model				WoW Test Unseen								
Widdei	BLEU-1	BLEU-4	RG-1	RG-L	METEOR	ACC	BLEU-1	BLEU-4	RG-1	RG-L	METEOR	ACC
TAKE(78%)	20.84	3.58	27.14	20.45	19.88	28.75	20.12	3.25	26.16	19.74	18.85	25.78
Random	19.58	2.95	25.84	19.40	18.57	23.60	18.59	2.64	24.70	18.67	17.35	18.04
80%	20.59	3.37	26.94	20.17	19.7	27.99	19.96	3.19	26.0	19.61	18.78	25.13
90%	20.98	3.65	27.37	20.61	20.03	29.97	20.3	3.32	26.37	19.93	19.04	27.34
Ideal(100%)	21.22	3.79	27.70	20.86	20.40	32.63	20.44	3.45	26.66	20.13	19.31	28.63

Table 7: More experiments on Wizard of Wikipedia. Random denotes that TAKE decides Sub-KS randomly. Ideal denotes that TAKE decides Sub-KS entirely depending on pseudo labels at the inference stage. 80% denotes that TAKE decides Sub-KS depending on 80% accurate pseudo labels.



Figure 3: Analysis of model with different pseudo label proportions on WOW. The red dots indicate experiment results under normal settings.

KGDG has three improvement directions: improving the accuracy of knowledge selection; improving the integration of external knowledge in generation (Zheng et al., 2021; Cui et al., 2021; Zhao et al., 2020b); improving the low-resource scenarios performance (Zheng and Huang, 2021; Zhao et al., 2020a; Liu et al., 2021). We mainly focus on the first direction. Lian et al. (2019) first proposed to utilize posterior knowledge to improve KS. Following (Lian et al., 2019), Kim et al. (2020) propose a sequential latent knowledge selection model; Chen et al. (2020b) attempt to bridge the gap between prior and posterior knowledge selection; Zhan et al. (2021a) find that sampling latent variable also helps response generation and proposed a collaborative latent variable model. Other work discovers more features in dialogue context to model KS (Zheng et al., 2020; Meng et al., 2020).

Topic-shift related works There is no general definition for "topic" (Purver et al., 2011). However, the definition given by Owen (1985) inspires our work in this paper. Although there have been various types of research about topic (Glavas and Somasundaran, 2020; Si et al., 2021), there is little work combining topic shift with multi-round dialogue. Xie et al. (2021) introduce a new topic-shift aware dialog benchmark TIAGE and three tasks. Sevegnani et al. (2021) propose a new dialogue connection task when the topic shifts. Zhan et al. (2021b) utilizes a BiLSTM-CRF network to predict topic tags before knowledge selection.

5 Conclusion

In this paper, we propose a Topic-shift Aware Knowledge sElector(TAKE) model, which better locates the relevant parts from dialogue history to improve the performance of knowledge selection. Besides, we obtain topic shift labels inspired by the idea of distant supervision and adopt curriculum learning and knowledge distillation methods to alleviate the negative influence of noises. Experiments on the WoW show that our model outperforms the baselines, and the ablation study indicates that all components of our methods work. In the future, we will research on multi-sentence knowledge integration further and combine the labeling work with knowledge graph.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 61976207, No. 61906187).

References

- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.
- Gillian R Brown and George Yule. 1983. Brown, gillian, and george yule, discourse analysis . cambridge, ma: Cambridge university press, 1983.
- Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. 2020. Bart for knowledge grounded conversations. In *Converse@KDD*.
- Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*, abs/1803.11175.
- Xiuyi Chen, Feilong Chen, Fandong Meng, Peng Li, and Jie Zhou. 2021. Unsupervised knowledge selection for dialogue generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1230–1244, Online. Association for Computational Linguistics.
- Xiuyi Chen, Fandong Meng, Peng Li, Feilong Chen, Shuang Xu, Bo Xu, and Jie Zhou. 2020a. Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3426–3437, Online. Association for Computational Linguistics.
- Xiuyi Chen, Fandong Meng, Peng Li, Feilong Chen, Shuang Xu, Bo Xu, and Jie Zhou. 2020b. Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3426–3437, Online. Association for Computational Linguistics.
- Leyang Cui, Yu Wu, Shujie Liu, and Yue Zhang. 2021. Knowledge enhanced fine-tuning for better handling unseen entities in dialogue generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2328–2337, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of

deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *CoRR*, abs/1811.01241.
- Mihail Eric, Nicole Chartier, Behnam Hedayatnia, Karthik Gopalakrishnan, Pankaj Rajan, Yang Liu, and Dilek Hakkani-Tur. 2021. Multi-sentence knowledge selection in open-domain dialogue. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 76–86, Aberdeen, Scotland, UK. Association for Computational Linguistics.
- Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.
- Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2017. A knowledge-grounded neural conversation model.
- Goran Glavas and Swapna Somasundaran. 2020. Twolevel transformer and auxiliary coherence modeling for improved text segmentation. In *AAAI*.
- Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.
- Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. *CoRR*, abs/2002.07510.
- Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.
- Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. *ArXiv*, abs/2107.07566.
- Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental transformer with deliberation decoder for document grounded conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 12–21, Florence, Italy. Association for Computational Linguistics.
- Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, pages 5081–5087. International Joint Conferences on Artificial Intelligence Organization.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

- Shilei Liu, Xiaofeng Zhao, Bochao Li, Feiliang Ren, Longhui Zhang, and Shujuan Yin. 2021. A Three-Stage Learning Framework for Low-Resource Knowledge-Grounded Dialogue Generation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2262– 2272, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Chuan Meng, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tengxiao Xi, and Maarten de Rijke. 2021. *Initiative-Aware Self-Supervised Learning for Knowledge-Grounded Conversations*, page 522–532. Association for Computing Machinery, New York, NY, USA.
- Chuan Meng, Pengjie Ren, Zhumin Chen, Weiwei Sun, Zhaochun Ren, Zhaopeng Tu, and Maarten de Rijke. 2020. DukeNet: A Dual Knowledge Interaction Network for Knowledge-Grounded Conversation, page 1151–1160. Association for Computing Machinery, New York, NY, USA.
- Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.
- Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2322–2332, Brussels, Belgium. Association for Computational Linguistics.
- Marion Owen. 1985. G. brown and g. yule, discourse analysis. cambridge: Cambridge university press, 1983. pp. xii 288. - m. stubbs, discourse analysis. oxford: Basil blackwell, 1983. pp. xiv 272. *Journal* of Linguistics, 21(1):241–245.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the* 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703.

- Matthew Purver, Arash Eshghi, and Julian Hough. 2011. Incremental semantic construction in a dialogue system. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011).*
- Karin Sevegnani, David M. Howcroft, Ioannis Konstas, and Verena Rieser. 2021. OTTers: One-turn topic transitions for open-domain dialogue. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 2492–2504, Online. Association for Computational Linguistics.
- Jiasheng Si, Deyu Zhou, Tongzhe Li, Xingyu Shi, and Yulan He. 2021. Topic-aware evidence reasoning and stance-aware aggregation for fact verification. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1612–1622, Online. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann A. Copestake. 2021. Tiage: A benchmark for topic-shift aware dialog modeling. *ArXiv*, abs/2109.04562.
- Haolan Zhan, Lei Shen, Hongshen Chen, and Hainan Zhang. 2021a. CoLV: A collaborative latent variable model for knowledge-grounded dialogue generation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2250–2261, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Haolan Zhan, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Yongjun Bao, and Yanyan Lan. 2021b. Augmenting knowledge-grounded conversations with sequential knowledge transition. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5621–5630, Online. Association for Computational Linguistics.
- Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020a. Low-resource knowledge-grounded dialogue generation. *ArXiv*, abs/2002.10348.

- Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020b. Knowledgegrounded dialogue generation with pre-trained language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3377–3390, Online. Association for Computational Linguistics.
- Chujie Zheng, Yunbo Cao, Daxin Jiang, and Minlie Huang. 2020. Difference-aware knowledge selection for knowledge-grounded conversation generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 115–125, Online. Association for Computational Linguistics.
- Chujie Zheng and Minlie Huang. 2021. Exploring prompt-based few-shot learning for grounded dialog generation. *CoRR*, abs/2109.06513.
- Wen Zheng, Natasa Milic-Frayling, and Ke Zhou. 2021. Knowledge-grounded dialogue generation with termlevel de-noising. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2972–2983, Online. Association for Computational Linguistics.
- Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.

A Example Appendix

A.1 Demonstrations of Pseudo Label

Table 8 shows a case on how we obtain pseudo labels.

A.2 More Cases

Example in Table 9 indicates that even though MIKe could select the golden knowledge like TAKE, TAKE has a better ability to integrate the ground source. MIKe tends to repeat the golden knowledge simply while TAKE generates more coherent and fluent responses.

	Topic: Canada
	(1) I have many friends in Canada, located in the northern part of North America.
Agent	Knowledge: <canada></canada> Canada is a country located in the northern part of North America.
	Pseudo label: 0
User	(2) I don't know anyone from Canada, what's it like?
	(3) It has 10 provinces.
	Knowledge: <canada> Its ten provinces and three territories extend from the Atlantic to</canada>
Agent	the Pacific and northward into the Arctic Ocean, covering, making it the world's
	second-largest country by total area.
	Pseudo label: 1
User	(4) In which provinces are some of your friends?
	(5) My friend lives in a place called Guelph. Canada is sparsely populated,
	most of its land is dominated by forest.
Agent	Knowledge: <canada> Canada is sparsely populated, the majority of its land territory being</canada>
	dominated by forest and tundra and the Rocky Mountains.
	Pseudo label: 1
User	(6) In Guelph, is it too cold?
	(7)It can be very cold there, especially in Wellington County.
Agont	Knowledge: <guelph> The area – much of what became Wellington County – had been</guelph>
Agem	part of the Halton Block, a Crown Reserve for the Six Nations Iroquois.
	Pseudo label: 0
User	(8) What are some interesting places in Wellington County?
	(9) Not many interesting places, the population there is only 90k.
Agont	Knowledge: < Wellington County, Ontario> According to the 2016 Census,
Agent	the population of the County was 90,932.
	Pseudo label: 0

Table 8: An example of labelling on the WoW dataset.

	Topic: Horse training
K_0	no passages used.
K_1	There is tremendous controversy over various methods of horse training and even some of the words used to describe these methods.
K_2	Historically, horses were trained for warfare, farm work, sport and transport.
K_3	Horses are trained to be manageable by humans for everyday care as well as for equestrian activities from horse racing to therapeutic horseback riding for people with disabilities.
K_4	Horse training refers to a variety of practices that teach horses to perform certain behaviors when asked to do so by humans.
context	(1)It amazes me that some can take a wild horse and train it to be so obedient. Do you know anything about horse training? (2)Yes! I'm a horse trainer! I use a variety of practices that teach horses to perform behaviors when asked. I love my job. Horses are such smart animals!(3) Please tell me more about what you do. I love horses and would love to know more.
DukeNet	$(K_2 \times)$ Well horses were trained for warfare, farm work, sport and transport.
KnowledGPT	$(K_2 \times)$ Well, I love horses, they were trained for warfare, farm work and transport.
MIKe	$(K_3\checkmark)$ Well, horses are trained to be manageable by humans for everyday care as well as therapeutic horseback riding.
TAKE	$(K_3\checkmark)$ I love to train them to be manageable by humans for everyday care. I also love to ride them for the rapeutic horseback riding.

Table 9: An example for case study. The utterances (2) to (4) exhibit a topic inheritance.