SEQUENTIAL LATENT KNOWLEDGE SELECTION FOR KNOWLEDGE-GROUNDED DIALOGUE

Anonymous authors
Paper under double-blind review

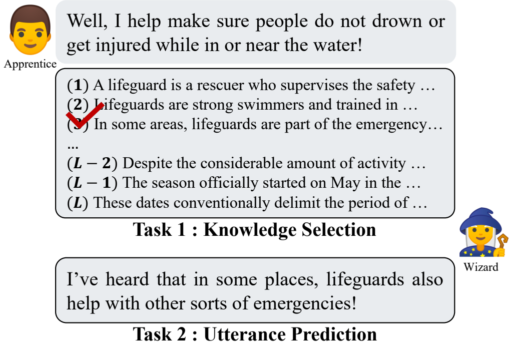
ABSTRACT

Knowledge-grounded dialogue is a task of generating an informative response based on both discourse context and external knowledge. As we focus on better modeling the *knowledge selection* in the multi-turn knowledge-grounded dialogue, we propose a sequential latent variable model as the first approach to this matter. The model named *sequential knowledge transformer* (SKT) can keep track of the prior and posterior distribution over knowledge; as a result, it can not only reduce the ambiguity caused from the diversity in knowledge selection of conversation but also better leverage the response information for proper choice of knowledge. Our experimental results show that the proposed model improves the knowledge selection accuracy and subsequently the performance of utterance generation. We achieve the new state-of-the-art performance on Wizard of Wikipedia (Dinan et al., 2019) as one of the most large-scale and challenging benchmarks. We further validate the effectiveness of our model over existing conversation methods in another knowledge-based dialogue Holl-E dataset (Moghe et al., 2018).

1 INTRODUCTION

Knowledge-grounded dialogue is a task of generating an informative response based on both discourse context and selected external knowledge (Ghazvininejad et al., 2018). For example, it is more descriptive and engaging to respond *"I've always been more of a fan of the American football team from Pittsburgh, the Steelers!"* than *"Nice, I like football too."*. As it has been one of the key milestone tasks in conversational research (Zhang et al., 2018), a majority of previous works have studied how to effectively combine given knowledge and dialogue context to generate an utterance (Zhang et al., 2018; Li et al., 2019b; Parthasarathi & Pineau, 2018; Madotto et al., 2018). Recently, Dinan et al. (2019) propose to tackle the knowledge-grounded dialogue by decomposing it into two sub-problems: first selecting knowledge from a large pool of candidates and generating a response based on the selected knowledge and context.

In this work, we investigate the issue of *knowledge selection* in the multi-turn knowledge-grounded dialogue, since practically the selection of pertinent topics is critical to better engage humans in conversation, and technically the utterance generation becomes easier with a more powerful and consistent knowledge selector in the system. Especially, we focus on developing a *sequential latent variable model* for knowledge selection, which has not been discussed in previous research. We believe it brings several advantages for more engaging and accurate knowledge-based chit-chat. First, it can correctly deal with the *diversity* in knowledge selection of conversation. Since one can choose any knowledge to carry on the conversation, there can be one-to-many relations between dialogue context and knowledge selection. Such multimodality by nature makes the training of a dialogue system much more difficult in a data-driven way. However, if we can sequentially model the history of knowledge selection in previous turns, we can reduce the scope of probable knowledge candidates at current turn. Second, the sequential latent model can better leverage the response information, which makes knowledge selection even more accurate. It is naturally easy to select the knowledge in the pool once the response is known, because the response is generated based on the selected knowledge. Our sequential model can keep track of prior and posterior distribution over knowledge, which are sequentially updated considering the responses in previous turns, and thus we can better predict the knowledge by sampling from the posterior. Third, the latent model works even when the knowledge selection labels for previous dialogue are not available, which is common



Well, I help make sure people do not drown or get injured while in or near the water!

(**1**) A lifeguard is a rescuer who supervises the safety …
(**2**) Lifeguards are strong swimmers and trained in …
(3) In some areas, lifeguards are part of the emergency…
...
($\boldsymbol{L-2}$) Despite the considerable amount of activity …
($\boldsymbol{L-1}$) The season officially started on May in the …
($\boldsymbol{L}$) These dates conventionally delimit the period of …

I've heard that in some places, lifeguards also help with other sorts of emergencies!

Figure 1: An example of wizard's tasks in knowledge-grounded conversation of Wizard of Wikipedia (Dinan et al., 2019).

Table 1: Accuracy of knowledge selection with and without knowing the response. We test with GRU (Cho et al., 2014), Transformer (Vaswani et al., 2017) and BERT (Devlin et al., 2019) as the sentence encoder. For human evaluation, we randomly sample 20 dialogues and ask human annotators to select the most likely knowledge sentence from the pool.

Methods	w/o response	w/ response
GRU	20.0	66.0
Transformer	22.5	70.4
BERT	23.4	78.2
Transformer + GT history	25.4	70.4
BERT + GT history	27.3	79.2
Random	2.7	2.7
Human	17.1	83.7

in practice. For example, if multiple people have discussion about given documents, knowledge selection of previous turns is done by others. The latent model can infer which knowledge others are likely to select and use.

Finally, the contributions of this work are as follows.

1. We propose a novel model named *sequential knowledge transformer* (SKT). To the best of our knowledge, our model is the first attempt to leverage a sequential latent variable model for knowledge selection, which subsequently improves the knowledge-grounded chit-chat.
2. Our experimental results show that the proposed model improves not only the knowledge selection accuracy but also the performance of utterance generation. As a result, we achieve the new state-of-the-art performance on Wizard of Wikipedia (Dinan et al., 2019) and a knowledge-annotated version of Holl-E (Moghe et al., 2018) dataset.

2 Problem Statement and Motivation

As a main testbed of our research, we choose the Wizard of Wikipedia (WoW) benchmark (Dinan et al., 2019), since it is one of the most large-scale and challenging datasets for open-domain multi-turn knowledge-based dialogue. Moreover, the dataset can evaluate the algorithm's ability for solving the two subproblems of knowledge selection and response generation. That is, it provides ground-truth labels of knowledge selection and clear grounding between the pairs of selected knowledge and response. In our experiments, we also evaluate on Holl-E (Moghe et al., 2018) as another dataset for knowledge-grounded dialogue, after collecting clearer labels of knowledge sentences.

The Flow of Conversation. The WoW (Dinan et al., 2019) deals with a chit-chat dialogue task where two speakers discuss in depth about a given topic. One speaker (coined as *Wizard*) is to be both engaging and knowledgeable on the topic with access to an information retrieval (IR) system over Wikipedia to supplement its knowledge. The other speaker (*Apprentice*) is curious and eager to learn about the topic. With an example in Figure 1, the conversation flow takes place as follows.

1. One topic is chosen among 1,431 topics and shared between the two speakers.
2. Given an apprentice's utterance and a wizard's previous utterance, the IR system retrieves relevant knowledge, which includes the first paragraph of top 7 articles each for wizard and apprentice and the first 10 sentences of the original Wikipedia page of the topic (*e.g.* the *lifeguard* wikipage). The knowledge pool contains 67.57 sentences on average. Then. the wizard must choose a single relevant sentence from them (*knowledge selection*) and construct an utterance (*response generation*).
3. The conversation repeats until a minimum number of turns (5 each) reaches.

The Motivation of Sequential Latent Models. The goal of the task is to model the wizard that solves the two subproblems of knowledge selection and response generation (Dinan et al., 2019). In the knowledge selection step, a single relevant knowledge sentence is chosen from a pool of candidates, and in the response generation step, a final utterance is generated with the chosen knowledge and dialogue context. This pipeline is originally proposed to tackle open-domain TextQA (Chen et al., 2017); for example, Min et al. (2018) show its effective for single-document TextQA, to which the key is to locate the sentences that contain the information about the answer to a question.

For knowledge-grounded dialogue, however, there can be one-to-many relations between the dialogue context and the knowledge to be selected unlike TextQA. Except a direct question about context, one can choose any diverse knowledge to carry on the conversation. Therefore, the knowledge selection in dialogue is diverse (*i.e.* multimodal) by nature, which should be correctly considered in the model. It is our main motivation to propose a sequential latent variable model for knowledge selection, which has not been studied yet. The latent variable not only models such diversity of knowledge but also sequentially track the topic flow of knowledge in the multi-turn dialogue.

Another practical advantage of the sequential latent model lies in that it is easy to find which knowledge is chosen once the response is known, since the response is written based on the selected knowledge. Table 1 clearly validates this relation between knowledge and response. In the WoW dataset, knowing a response boosts the accuracy of knowledge sentence selection for both human and different models. These results hint that knowledge selection may need to be jointly modeled with response generation in a sequence of multi-turn chit-chats, which can be done by the sequential latent models.

3 Approach

We propose a novel model for knowledge-grounded conversation named *sequential knowledge transformer* (SKT), whose graphical model is illustrated in Figure 2. It is a sequential latent model that sequentially conditions on previously selected knowledge to generate a response.

We will use $1 \leq t \leq T$ to iterate over dialogue turns, $1 \leq m \leq M$ and $1 \leq n \leq N$ to respectively iterate over words in the utterance of apprentice and wizard, and $1 \leq l \leq L$ to denote knowledge sentences in the pool. Thus, T is the dialogue length, M and N are the length of each utterance of apprentice and wizard, and L is the size of the knowledge pool.

The input to our model at turn t is previous turns of conversation, which consists of utterances from apprentice $\mathbf{x}^1, ..., \mathbf{x}^t$, utterances from wizard $\mathbf{y}^1, ..., \mathbf{y}^{t-1}$ and the knowledge pool $\mathbf{k}^t = \{\mathbf{k}^{t,l}\} = \mathbf{k}^{t,1}, ..., \mathbf{k}^{t,L}$. The output of the model is selected knowledge $\mathbf{k}_s^t$ and the wizard's response $\mathbf{y}^t$. Below, we discuss sentence embedding, knowledge selection and utterance decoding in our approach. Note that our technical novelty lies in the knowledge selection model, while exploiting existing techniques for text encoding and utterance decoding.

Sentence Encoding. We represent an apprentice utterance $\mathbf{x}^t$ to an embedding $\mathbf{h}_x^t$ using BERT (Devlin et al., 2019) and average pooling over time steps (Cer et al., 2018):

$$\mathbf{H}_x^t = \text{BERT}_{base}([x_1^t; ...; x_M^t]) \in \mathbb{R}^{M \times 768}, \mathbf{h}_x^t = \text{avgpool}(\mathbf{H}_x^t) \in \mathbb{R}^{768}. \quad (1)$$

Likewise, the utterance of Wizard $\mathbf{y}^{t-1}$ is embedded as $\mathbf{h}_y^{t-1}$ and knowledge sentences are as $\{\mathbf{h}_k^{t,l}\} = \mathbf{h}_k^{t,1}, ..., \mathbf{h}_k^{t,L}$. Each apprentice-wizard utterance pair $\mathbf{h}_{xy}^t = [\mathbf{h}_x^t; \mathbf{h}_y^t]$ at dialog turn t is jointly represented through a GRU (Cho et al., 2014) layer: $\mathbf{d}_{xy}^t = \text{GRU}_{dialog}(\mathbf{d}_{xy}^{t-1}, \mathbf{h}_{xy}^t) \in \mathbb{R}^{768}$.

Sequential Knowledge Selection. Compared to previous works, we make two significant modifications. First, we regard the knowledge selection as a sequential decision process instead of a single-step decision process. Second, due to the diversity of knowledge selection in dialogue, we model it as latent variables. As a result, we can carry out the joint inference of multi-turns of knowledge selection and response generation rather than separate inference turn by turn.

There have been much research on sequential latent variable models (Chung et al., 2015; Fraccaro et al., 2016; Goyal et al., 2017; Aneja et al., 2019; Shankar & Sarawagi, 2019). For example, Shankar & Sarawagi (2019) propose a posterior attention model that represents the attention of seq2seq models as sequential latent variables. Inspired by them, we factorize the response generation with

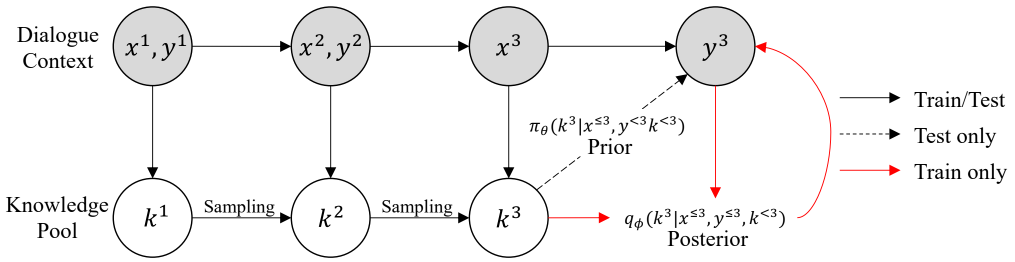


Figure 2: A graphical representation of the proposed *sequential knowledge selection* (SKT) model. At the third turn, the goal is to generate wizard's response ($\mathbf{y}^3$) given dialogue context ($\mathbf{x}^{\leq 3}, \mathbf{y}^{<3}$). Our model sequentially infer which knowledge is likely to be used ($\mathbf{k}^{\leq 3}$), from which the utterance $\mathbf{y}^3$ is generated.

latent knowledge selection and derive the variational lower bound as follows:

$$\log p(\mathbf{y}|\mathbf{x}) = \log \prod_t \sum_{\mathbf{k}^t} p_\theta(\mathbf{y}^t|\mathbf{x}^{\leq t}, \mathbf{y}^{<t}, \mathbf{k}^{\leq t}) \pi_\theta(\mathbf{k}^t|\mathbf{x}^{\leq t}, \mathbf{y}^{<t}, \mathbf{k}^{<t}) \tag{2}$$

$$\geq \sum_t \mathbb{E}_{q_\phi(\mathbf{k}^{t-1})} \Big[\mathbb{E}_{q_\phi(\mathbf{k}^t)} [\log p_\theta(\mathbf{y}^t|\mathbf{x}^{\leq t}, \mathbf{y}^{<t}, \mathbf{k}^t)] - D_{KL}(q_\phi(\mathbf{k}^t) \parallel \pi_\theta(\mathbf{k}^t)) \Big], \tag{3}$$

where $q_\phi(\mathbf{k}^t)$ is shorthand for $q_\phi(\mathbf{k}^t|\mathbf{x}^{\leq t}, \mathbf{y}^{\leq t}, \mathbf{k}^{<t})$ and $\pi_\theta(\mathbf{k}^t)$ for $\pi_\theta(\mathbf{k}^t|\mathbf{x}^{\leq t}, \mathbf{y}^{<t}, \mathbf{k}^{<t})$ for brevity. Note that $p_\theta(\mathbf{y}^t|\cdot)$ is a decoder network, $\pi_\theta(\mathbf{k}^t)$ is a categorical conditional distribution of knowledge given dialogue context and previously selected knowledge, and $q_\phi(\mathbf{k}^t)$ is an inference network to approximate posterior distribution $p_\theta(\mathbf{k}^t|\mathbf{x}^{\leq t}, \mathbf{y}^{\leq t}, \mathbf{k}^{<t})$.

The conditional probability of generating wizard's response $\mathbf{y}^t$ given dialogue context $\mathbf{x}^{\leq t}$ and $\mathbf{y}^{<t}$, can be re-written from Eq. (2) as follows:

$$p(\mathbf{y}^t|\mathbf{x}^{\leq t}, \mathbf{y}^{<t}) \approx \prod_{i=1}^{t-1} \sum_{\mathbf{k}^i} q_\phi(\mathbf{k}^i) \Big(\sum_{\mathbf{k}^t} p_\theta(\mathbf{y}^t|\mathbf{x}^{\leq t}, \mathbf{y}^{<t}, \mathbf{k}^t) \pi_\theta(\mathbf{k}^t) \Big). \tag{4}$$

The detailed derivation can be found in Appendix. Eq.(4) means that we first infer from the knowledge posterior which knowledge would be used up to previous turn $t-1$, estimate the knowledge for current turn t from prior knowledge distribution and generate an utterance from the inferred knowledge. Figure 2 shows an example of this generation process at $t = 3$. We parameterize the decoder network p_θ, the prior distribution of knowledge π_θ, and the approximate posterior q_ϕ with deep neural networks as will be discussed.

From the posterior distribution $q_\phi(\mathbf{k}^{t-1})$ we draw a sample $\mathbf{k}_s^{t-1}$, and then update π_θ and q_ϕ with the sentence embedding of sampled knowledge ($\mathbf{h}_k^{t-1,s}$) and the embeddings of previous and current utterances ($\mathbf{d}_{xy}^{t-1}, \mathbf{d}_{xy}^t, \mathbf{h}_x^t$). We use an attention mechanism over current knowledge pool $\{\mathbf{h}_k^{t,l}\}$ to compute knowledge distribution given the dialogue context. This process is modeled as

$$\pi_\theta(\mathbf{k}^t|\mathbf{x}^{\leq t}, \mathbf{y}^{<t}, \mathbf{k}_s^{\leq t-1}) = \text{softmax}(\mathbf{q}_{prior}^t [\mathbf{h}_k^{t,1}, ..., \mathbf{h}_k^{t,L}]^\top) \in \mathbb{R}^L \tag{5}$$

$$q_\phi(\mathbf{k}^t|\mathbf{x}^{\leq t}, \mathbf{y}^{\leq t}, \mathbf{k}_s^{\leq t-1}) = \text{softmax}(\mathbf{q}_{post}^t [\mathbf{h}_k^{t,1}, ..., \mathbf{h}_k^{t,L}]^\top) \in \mathbb{R}^L, \tag{6}$$

where

$$\mathbf{q}_{prior}^t = \mathbf{W}_{prior}([\mathbf{d}_{xy}^{t-1}; \mathbf{h}_x^t; \text{GRU}_{hist}(\mathbf{d}_k^{t-2}, \mathbf{h}_k^{t-1,s})]), \tag{7}$$

$$\mathbf{q}_{post}^t = \mathbf{W}_{post}([\mathbf{d}_{xy}^t; \text{GRU}_{hist}(\mathbf{d}_k^{t-2}, \mathbf{h}_k^{t-1,s})]), \tag{8}$$

$\mathbf{d}_k^t$ is the hidden state of GRU_{hist} and we initialize $\mathbf{d}_{xy}^0 = \mathbf{d}_k^0 = \mathbf{0} \in \mathbb{R}^{768}$, and $\mathbf{W}_{prior}, \mathbf{W}_{post} \in \mathbb{R}^{768 \times (768 \times 2)}$ are the parameters. We here use the GRU (Li et al., 2017; Aneja et al., 2019) to sequentially condition previously selected knowledge to π_θ and q_ϕ.

Finally, we sample knowledge $\mathbf{k}_s^t$ over attention distribution in Eq.(5) and pass it to the decoder. At test time, we select the knowledge with the highest probability.

Decoding with Copy Mechanism. We generate the wizard's response at turn t, given current context $\mathbf{x}^t$ and selected knowledge sentence $\mathbf{k}_s^t$. We feed their concatenated embedding

$\mathbf{H}^t_{xk_s} = [\mathbf{H}^t_x; \mathbf{H}^t_{k_s}]$ to the decoder p_θ. To maximize the effect of selected knowledge for response generation, we choose the Copy mechanism (Xia et al., 2017; Li et al., 2019b) with Transformer decoder (Vaswani et al., 2017). We obtain the output word probability (Zhao et al., 2019a):

$$\mathbf{h}^t_n = \text{Decoder}(\mathbf{H}^t_{xk_s}, \mathbf{y}^t_{<n}), \quad \mathbf{q}^t_n, \mathbf{K}^t, \mathbf{V}^t = \mathbf{h}^t_n \mathbf{W}^\top_q, \mathbf{H}^t_{xk_s} \mathbf{W}^\top_k, \mathbf{H}^t_{xk_s} \mathbf{W}^\top_v, \tag{9}$$

$$p^{gen}_{t,n}(w) = \text{softmax}(\mathbf{W}_{out} \mathbf{h}^t_n), \quad p^{copy}_{t,n}(w) = \text{softmax}(\mathbf{q}^t_n \mathbf{K}^t), \tag{10}$$

$$p_{t,n}(w) = (1 - \alpha^{copy}_{t,n}) * p^{gen}_{t,n}(w) + \alpha^{copy}_{t,n} * p^{copy}_{t,n}(w), \tag{11}$$

where $\alpha^{copy}_{t,n} = \sigma(\mathbf{W}^\top_{copy} \sum p^{copy}_{t,n}(w) \cdot \mathbf{V}^t)$ and σ is a sigmoid. Finally, we select the word with the highest probability $y^t_{n+1} = \arg\max_{w \in \mathcal{V}} p_{t,n}(w)$ where $\mathcal{V}$ is the dictionary. Unless the word y^t_{n+1} is an EOS token, we repeat generating the next word by feeding y^t_{n+1} to the decoder.

3.1 Training

Obviously, there is a large gap in knowledge selection accuracy between training with or without true labels (*e.g.* 23.2 of E2E Transformer MemNet with labels vs 4.8 of PostKS without labels in Table 2). As one way to take advantage of true labels for training of latent models, prior research has employed auxiliary losses over latent variables (Wen et al., 2017; Zhao et al., 2017). Similarly, we use the knowledge loss from Dinan et al. (2019) (*i.e.* the cross-entropy loss between predicted and true knowledge sentences) as an auxiliary loss for the latent variable. Thus, the training objective is a combination of the variational lower-bound from Eq. (3) and the auxiliary knowledge loss as

$$\mathcal{L} = -\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{q_\phi(\mathbf{k}^{t-1})} \Big[\mathbb{E}_{q_\phi(\mathbf{k}^t)} [\log p_\theta(\mathbf{y}^t | \mathbf{x}^{\leq t}, \mathbf{y}^{<t}, \mathbf{k}^t_s)] \\ - D_{KL}(q_\phi(\mathbf{k}^t) \parallel \pi_\theta(\mathbf{k}^t)) + \lambda \underbrace{\log q_\phi(\mathbf{k}^t_a)}_{\text{Knowledge loss}} \Big], \tag{12}$$

where $\mathbf{k}^t_s$ is a sampled knowledge from $q_\phi(\mathbf{k}^t | \mathbf{x}^{\leq t}, \mathbf{y}^{\leq t}, \mathbf{k}^{<t})$, $\mathbf{k}^t_a$ is a true knowledge, and λ is a hyperparameter. Note that knowledge is sequentially sampled from attention distribution as in Eq.(5). We train our model by mini-batch gradient descent. We approximate the expectation of one sample from the posterior by using Gumbel-Softmax function (Jang et al., 2017; Maddison et al., 2017b). Further details of optimization can be found in Appendix.

4 Experiments

We evaluate our model mainly on the Wizard of Wikipedia (Dinan et al., 2019) and additionally Holl-E (Moghe et al., 2018) as another knowledge-grounded chit-chat dataset. We qualitatively and quantitatively compare our approach with other state-of-the-art models.

4.1 Datasets

Wizard of Wikipedia. It contains 18,430 dialogues for training, 1,948 dialogues for validation and 1,933 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.

Holl-E. It contains 7,228 dialogues for training, 930 dialogues for validation and 913 dialogues for test. A single document is given per dialogue; the documents include about 58 and 63 sentences on average for training/validation and test set, respectively. The dataset provides *spans* in the document as additional information to provide which parts of a document is used to generate a response. However, the span labels are highly inconsistent; for example, they are often shorter than a single sentence or contain multiple consecutive sentences. Thus, it is undesirable to use them without modifications because it is different from WoW setting where all of the ground-truth (GT) knowledge are in the form of sentence. Hence, we collect a set of ground-truth (GT) knowledge as follows. If the span is given as multiple sentences, we select the closest sentence in the span to the response in terms of F1 scores. Otherwise, we select the closest sentence in the whole document. If all sentences have zero F1 scores to the response, we tag *no passages used* as the GT, which amounts to 5% of GT labels. We will make our set of GT annotations public.

Table 2: Quantitative results on the Wizard of Wikipedia dataset (Dinan et al., 2019). The method with [*] does not use the knowledge loss. The scores of E2E Transformer MemNet† and Transformer (no knowledge)† are from the original paper. The variant (BERT vocab)‡ is re-runned using the authors' code, since the vocabulary is different from original paper due to the use of BERT.

	Test Seen				**Test Unseen**			
Method	PPL	R-1	R-2	Acc	PPL	R-1	R-2	Acc
Random knowledge selection	n/a	8.4	1.4	2.7	n/a	8.0	1.2	2.3
Repeat last utterance	n/a	14.5	3.1	n/a	n/a	14.1	2.9	n/a
Transformer (no knowledge)† (Dinan et al., 2019)	**41.8**	17.8	n/a	n/a	87.0	14.0	n/a	n/a
E2E Transformer MemNet† (Dinan et al., 2019)	63.5	16.9	n/a	22.5	97.3	14.4	n/a	12.2
E2E Transformer MemNet (BERT vocab)‡	53.2	17.7	4.8	23.2	137.8	13.6	1.9	10.5
PostKS* (Lian et al., 2019)	79.1	13.0	1.0	4.8	193.8	13.1	1.0	4.2
E2E Transformer MemNet + BERT	53.5	16.8	4.5	23.7	105.7	13.5	2.2	13.6
PostKS + Knowledge Loss	54.5	18.1	5.3	23.4	144.8	13.5	2.0	9.4
E2E Transformer MemNet + BERT + PostKS	54.6	17.8	5.3	25.5	113.2	13.4	2.3	14.1
Ours	52.1	**19.3**	**6.7**	**26.8**	**79.9**	**16.2**	**4.2**	**19.2**

Table 3: Quantitative results on the Holl-E dataset (Moghe et al., 2018) with single reference and multiple references test set.

	Single Reference				**Multiple References**			
Method	PPL	R-1	R-2	Acc	PPL	R-1	R-2	Acc
Random knowledge selection	n/a	7.4	1.8	1.9	n/a	10.3	3.6	3.5
Repeat last utterance	n/a	11.4	1.5	n/a	n/a	13.6	2.0	n/a
E2E Transformer MemNet (Dinan et al., 2019)	140.6	20.1	10.3	22.7	83.6	24.3	12.8	32.3
PostKS* (Lian et al., 2019)	196.6	15.2	6.0	1.5	114.1	19.2	7.9	3.2
E2E Transformer MemNet + BERT	112.6	25.9	18.3	28.2	66.9	31.1	22.7	37.5
PostKS + Knowledge Loss	135.1	19.9	10.7	22.5	81.9	23.8	12.9	32.2
E2E Transformer MemNet + BERT + PostKS	119.9	27.8	20.1	27.6	66.7	33.7	25.8	37.3
Ours	**51.8**	**29.5**	**23.0**	**30.3**	**29.6**	**36.4**	**29.7**	**40.5**

4.2 Experimental Setting

Evaluation Metrics. We follow the evaluation protocol of WoW (Dinan et al., 2019). We measure unigram F1 (R-1), bigram F1 (R-2) and perplexity (PPL) for response generation, and the accuracy for knowledge selection. For n-gram metrics, we remove all the punctuations and (a, an, the) before computing the score. We remind that lower perplexity and higher n-gram (R-1, R-2) scores indicate better performance.

The test set for Holl-E is split into two subsets, *single reference* and *multiple references*. The dataset basically provides a single response per context (denoted as single reference). However, for some conversations, more responses (*e.g.* 2–13) are collected from multiple annotators per context (multiple references). For evaluation of multiple references, we take the best score over multiple GTs by following Moghe et al. (2018). For knowledge accuracy, we regard the model's prediction is correct if it matches at least one of the correct answers.

Baselines. We closely compare with two state-of-the-art knowledge-grounded dialogue models. The first one is E2E Transformer MemNet (Dinan et al., 2019), which uses a Transformer memory network for knowledge selection and a Transformer decoder for utterance prediction. The second one is PostKS (Lian et al., 2019), which uses the posterior knowledge distribution as a pseudo-label for knowledge selection. For fair comparison, we replace all GRU layers in PostKS with Transformers. We also compare with three variants of these models as an ablation study: (i) E2E Transformer MemNet + BERT, where we replace the Transformer memory network with pre-trained BERT, (ii) PostKS + Knowledge loss, where we additionally use the knowledge loss and (iii) E2E Transformer MemNet + BERT + PostKS, which combines all the components of baselines. We use official BERT tokenizer to tokenize the words and use pre-defined BERT vocabulary ($\mathcal{V} = 30522$) to convert token to index[1]. All the baselines use the exactly same inputs with our model except PostKS, which does not make use of knowledge labels as proposed in the original paper.

[1] `https://github.com/tensorflow/models/tree/master/official/nlp/bert`.

Table 4: Human evaluation results on the Wizard of Wikipedia. We report the mean ratings and their standard errors of different methods for engagingness and knowledgeability scores.

	Test Seen		**Test Unseen**	
Method	Engagingness	Knowledgeability	Engagingness	Knowledgeability
PostKS	1.65 (0.05)	1.72 (0.06)	1.66 (0.06)	1.74 (0.06)
E2E Transformer MemNet	2.57 (0.05)	2.47 (0.06)	2.39 (0.06)	2.21 (0.06)
Ours	2.59 (0.05)	2.53 (0.06)	2.52 (0.06)	2.35 (0.06)
Human	3.14 (0.05)	3.09 (0.05)	3.11 (0.05)	2.99 (0.05)

4.3 Quantitative Results

Table 2 compares the performance of different methods on the Wizard of Wikipedia dataset. Our model outperforms the state-of-the-art knowledge-grounded dialogue models in all metrics for knowledge selection (accuracy) and utterance generation (unigram F1, bigram F1). The PostKS that is trained with no knowledge label shows low accuracy on knowledge selection, which is slightly better than random guess. However, it attains better performance than E2E Transformer MemNet with the knowledge loss in the WoW Test Seen, which shows that leveraging prior and posterior knowledge distribution is effective for knowledge-grounded dialogue, although using sequential latent variable improves further. BERT improves knowledge selection accuracy, but not much as in TextQA because of diversity in knowledge selection of conversation. The E2E Transformer MemNet + BERT + PostKS performs the best among baselines, but not as good as ours, which validates that sequential latent modeling is critical for improving the accuracy of knowledge selection and subsequently utterance generation. Additionally, the performance gaps between ours and baselines are larger in Test Unseen. It can be understood that the sequential latent variable can generalize better. Transformer (no knowledge) shows the lowest perplexity in the WoW Test Seen, and it is mainly due to that it may generate only general and simple utterances since no knowledge is grounded. This behavior can be advantageous for the perplexity, while the other knowledge-based models take a risk of predicting wrong knowledge, which is unfavorable for perplexity.

Table 3 compares the performance of our model on Holl-E dataset. Similarly, our model outperforms all the baselines in all metrics. One notable trend is that BERT considerably reduces the perplexity in all models, which may be due to that the dataset size of Holl-E is much smaller than WoW and BERT prevents overfitting (Hao et al., 2019).

4.4 Qualitative Results

User Studies. We perform human evaluation to complement the limitation of automatic language metrics. We evaluate several aspects of utterance generation using the similar setting in Guu et al. (2018). We randomly sample 100 test examples, and each sample is evaluated by three unique human annotators on Amazon Mechanical Turk (AMT). At test, we show dialogue context and generated utterance by our method or baselines. We ask turkers to rate the quality of each utterance in two aspects, which are referred to Li et al. (2019a): (i) *Engagingness*: how much do you like the response? and (ii) *Knowledgeability*: how much is the response informative? Each item is scored from 1 to 4, where 1 means *not at all*, 2 is *a little*, 3 is *somewhat*, and 4 is *a lot*. Note that human evaluation on knowledge selection is not possible, since any knowledge can be fine for a given context, which is key motivation for our sequential latent model – *diversity* of knowledge selection.

Table 4 summarizes the results of the human evaluation, which validates that annotators prefer our results to those of baselines. Again, the performance gaps between ours and baselines are larger in Test Unseen, thank to better generality of our sequential latent model. However, the gaps in Test Seen are not large, since the evaluation is not done in a multi-turn setting, for which our sequential model's merit would be more salient, due to the difficulty of multi-turn tasks for AMT turkers.

Dialogue Examples. Figure 3 shows selected examples of utterance prediction. In each set, we show dialogue context, human response, and utterances generated by our method and baselines. Thanks to the use of latent variables, our model can better capture the changes in dialogue topics and thus generate more appropriate responses.

Seen Test (Topic: Italian Cuisine)	**Unseen Test (Topic: Hunting)**
⋮	⋮
A: I love chicken parmigiana as well, but I think my ultimate favorite is beef lasagna… Extra cheese please! W: Chicken with sauce and mozzarella…. Be still my heart! A: Truthfully, anything with cheese is the best	W: That is true but we always have to watch out for excessive hunting. It has caused some species to be endangered. A: Yes I agree. I don't believe in the useless hunting that poachers do. Its so cruel.
(Ours) i love pizza too ! it ' s a traditional italian dish consisting of yeasted flatbread typically topped with tomato sauce and cheese **(TMN)** i love cheese ! **(TMN+KL+BERT)** i like mine topped with vegetables , meats , and condiments . **(Human)** especially cheddar cheese ! it ' s the second most popular cheese in the use !	**(Ours)** i agree , poaching has been defined as the illegal hunting or capturing of wild animals . **(TMN)** i thinks so , i ' m not sure if you ' re talking about poaching , but i know that poodles are the second most intelligent breed behind the poodle . **(TMN+KL+BERT)** i agree . i think it ' s a great way to catch fish . **(Human)** agreed , i remember reading one time that unless you plan to kill the animals its not considered hunting .

Figure 3: Examples of generated responses by our model and baselines on Wizard of Wikipedia. TMN stands for E2E Transformer MemNet, and A and W for apprentice and wizard.

5 Related Work

Knowledge-based conversations have been studied much including collecting new datasets (Qin et al., 2019; Zhang et al., 2018; Ghazvininejad et al., 2018; Zhou et al., 2018; Dinan et al., 2019; Moghe et al., 2018) or developing new models (Lian et al., 2019; Li et al., 2019b; Yavuz et al., 2019; Zhao et al., 2019b; Dinan et al., 2019; Liu et al., 2019). Most works on the models have less investigated the knowledge selection issue but instead focused on how to effectively combine given knowledge and dialogue context to improve response informativeness. For example, Ghazvininejad et al. (2018) aid a Seq2Seq model with an external knowledge memory network, and Li et al. (2019b) propose an Incremental Transformer to encode multi-turn utterances along with knowledge in related documents. Recently, Dinan et al. (2019) propose both a dataset of Wizard of Wikipedia and a model to leverage the two-step procedure of selecting knowledge from the pool and generating a response based on chosen knowledge and given context.

One of the most related models to ours may be Lian et al. (2019), who also focus on the knowledge selection issue in the two-stage knowledge-grounded dialogue. However, our work is novel in that we model it as a sequence decision process with latent variables and introduce the knowledge loss. Thanks to these updates, our model achieves significantly better performance as shown in the experiments.

Sequential Latent Variable Models. There have been many studies about sequential latent variable models. Chung et al. (2015) propose one of the earliest latent models for sequential data, named VRNN. Later, this architecture is extended to SRNN (Fraccaro et al., 2016) and Z-Forcing (Goyal et al., 2017). There have been some notable applications of sequential latent models, including document summarization (Li et al., 2017), image captioning (Aneja et al., 2019) and text generation (Shao et al., 2019). Another related class of sequential latent models may be *latent attention models* (Deng et al., 2018; Wang et al., 2018; Yang et al., 2017), which exploit the latent variables to model the attention mapping between input and output sequences. Although our method is partly influenced by such recent models, it is novel to propose a sequential latent model for the knowledge-grounded chit-chat problem.

6 Conclusion

This work investigated the issue of knowledge selection in multi-turn knowledge-grounded dialogue, and proposed a sequential latent variable model, for the first time, named *sequential knowledge transformer* (SKT). Our method achieved the new state-of-the-art performance on the Wizard of Wikipedia benchmark (Dinan et al., 2019) and a knowledge-annotated version of Holl-E dataset (Moghe et al., 2018). There are several promising future directions beyond this work. First, we can explore other inference models such as sequential Monte Carlo methods using *filtering variational objectives* (Maddison et al., 2017a). Second, we can study the interpretability of knowledge selection such as measuring the uncertainty of attention (Heo et al., 2018).

REFERENCES

Jyoti Aneja, Harsh Agrawai, Dhruv Batra, and Alexander Schwing. Sequential Latent Spaces for Modeling the Intention During Diverse Image Captioning. In *ICCV*, 2019.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. In *TACL*, 2016.

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. Universal Sentence Encoder. *arXiv:1803.11175*, 2018.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to Answer Open-Domain Questions. In *ACL*, 2017.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. In *EMNLP*, 2014.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A Recurrent Latent Variable Model for Sequential Data. In *NIPS*, 2015.

Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, and Alexander Rush. Latent Alignment and Variational Attention. In *NIPS*, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, 2019.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *ICLR*, 2019.

Sergey Edunov, Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. Classical Structured Prediction Losses for Sequence to Sequence Learning. In *NAACL-HLT*, 2017.

Marco Fraccaro, Søren Kaae Sønderby, Ulrich Paquet, and Ole Winther. Sequential Neural Models with Stochastic Layers. In *NIPS*, 2016.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. A Knowledge-Grounded Neural Conversation Model. In *AAAI*, 2018.

Xavier Glorot and Yoshua Bengio. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *AISTATS*, 2010.

Anirudh Goyal Alias Parth Goyal, Alessandro Sordoni, Marc-Alexandre Côté, Nan Rosemary Ke, and Yoshua Bengio. Z-Forcing: Training Stochastic Recurrent Networks. In *NIPS*, 2017.

Kelvin Guu, Tatsunori B Hashimoto, Yonatan Oren, and Percy Liang. Generating Sentences by Editing Prototypes. *TACL*, 6:437–450, 2018.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Visualizing and Understanding the Effectiveness of BERT. In *EMNLP*, 2019.

Jay Heo, Hae Beom Lee, Saehoon Kim, Juho Lee, Kwang Joon Kim, Eunho Yang, and Sung Ju Hwang. Uncertainty-Aware Attention for Reliable Interpretation and Prediction. In *NIPS*, 2018.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical Reparameterization with Gumbel-Softmax. In *ICLR*, 2017.

Diederik Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015.

Margaret Li, Jason Weston, and Stephen Roller. ACUTE-EVAL: Improved Dialogue Evaluation with Optimized Questions and Multi-turn Comparisons. *arXiv:1909.03087*, 2019a.

Piji Li, Wai Lam, Lidong Bing, and Zihao Wang. Deep Recurrent Generative Decoder for Abstractive Text Summarization. In *EMNLP*, 2017.

Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. Incremental Transformer with Deliberation Decoder for Document Grounded Conversations. In *ACL*, 2019b.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. Learning to Select Knowledge for Response Generation in Dialog Systems. In *IJCAI*, 2019.

Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. Knowledge aware conversation generation with reasoning on augmented graph. In *EMNLP*, 2019.

Chris J Maddison, John Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, and Yee Teh. Filtering Variational Objectives. In *NIPS*, 2017a.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. 2017b.

Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. Mem2Seq: Effectively Incorporating Knowledge Bases into End-to-End Task-Oriented Dialog Systems. In *EMNLP*, 2018.

Sewon Min, Victor Zhong, Richard Socher, and Caiming Xiong. Efficient and Robust Question Answering from Minimal Context over Documents. In *ACL*, 2018.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M Khapra. Towards Exploiting Background Knowledge for Building Conversation Systems. In *EMNLP*, 2018.

Prasanna Parthasarathi and Joelle Pineau. Extending Neural Generative Conversational Model using External Knowledge Sources. In *EMNLP*, 2018.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing Neural Networks by Penalizing Confident Output Distributions. In *ICLR*, 2017.

Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. Conversing by Reading: Contentful Neural Conversation with On-Demand Machine Reading. In *ACL*, 2019.

Shiv Shankar and Sunita Sarawagi. Posterior Attention Models for Sequence to Sequence Learning. In *ICLR*, 2019.

Zhihong Shao, Minlie Huang, Jiangtao Wen, Wenfei Xu, and Xiaoyan Zhu. Long and Diverse Text Generation with Planning-based Hierarchical Variational Model. In *EMNLP*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In *NIPS*, 2017.

Weiyue Wang, Derui Zhu, Tamer Alkhouli, Zixuan Gan, and Hermann Ney. Neural Hidden Markov Model for Machine Translation. In *ACL*, 2018.

Tsung-Hsien Wen, Yishu Miao, Phil Blunsom, and Steve Young. Latent intention dialogue models. In *ICML*, 2017.

Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. Deliberation Networks: Sequence Generation Beyond One-Pass Decoding. In *NIPS*, 2017.

Zichao Yang, Zhiting Hu, Yuntian Deng, Chris Dyer, and Alex Smola. Neural Machine Translation with Recurrent Attention Modeling. In *EACL*, 2017.

Semih Yavuz, Abhinav Rastogi, Guan-Lin Chao, and Hakkani-Tur Dilek. DeepCopy: Grounded Response Generation with Hierarchical Pointer Networks. In *SIGDIAL*, 2019.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing Dialogue Agents: I Have a Dog, Do You Have Pets Too? In *ACL*, 2018.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In *ACL*, 2017.

Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. Improving Grammatical Error Correction via Pre-Training a Copy-Augmented Architecture with Unlabeled Data. In *NAACL-HLT*, 2019a.

Xueliang Zhao, Chongyang Tao, Wei Wu, Can Xu, Dongyan Zhao, and Rui Yan. A Document-Grounded Matching Network for Response Selection in Retrieval-based Chatbots. In *IJCAI*, 2019b.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. A Dataset for Document Grounded Conversations. In *EMNLP*, 2018.

A Derivation of Conditional Probability

In Section 3, we re-write the conditional probability of wizard's response $\mathbf{y}^t$ given dialogue context $\mathbf{x}^{\leq t}$ and $\mathbf{y}^{<t}$ from Eq. (2) to Eq. (4). We can simply derive it as follows:

$$
\begin{aligned}
& p(\mathbf{y}|\mathbf{x}) && (13)\\
&= \prod_{t} \sum_{\mathbf{k}^t} p_\theta(\mathbf{y}^t|\mathbf{x}^{\leq t}, \mathbf{y}^{<t}, \mathbf{k}^{\leq t})\pi_\theta(\mathbf{k}^t) \quad \text{(by Eq. (2))} && (14)\\
&= \prod_{i=1}^{t-1} \sum_{\mathbf{k}^i} p_\theta(\mathbf{y}^i|\mathbf{x}^{\leq i}, \mathbf{y}^{<i}) p_\theta(\mathbf{k}^i) \Big(\sum_{\mathbf{k}^t} p_\theta(\mathbf{y}^t|\mathbf{x}^{\leq t}, \mathbf{y}^{<t}, \mathbf{k}^{\leq t})\pi_\theta(\mathbf{k}^t) \Big) \quad \text{(by Bayes' rule)} && (15)\\
&\approx \prod_{i=1}^{t-1} \sum_{\mathbf{k}^i} p_\theta(\mathbf{y}^i|\mathbf{x}^{\leq i}, \mathbf{y}^{<i}) q_\phi(\mathbf{k}^i) \Big(\sum_{\mathbf{k}^t} p_\theta(\mathbf{y}^t|\mathbf{x}^{\leq t}, \mathbf{y}^{<t}, \mathbf{k}^{\leq t})\pi_\theta(\mathbf{k}^t) \Big) && (16)\\
&= \prod_{i=1}^{t-1} \sum_{\mathbf{k}^i} q_\phi(\mathbf{k}^i) \Big(\sum_{\mathbf{k}^t} p_\theta(\mathbf{y}^t|\mathbf{x}^{\leq t}, \mathbf{y}^{<t}, \mathbf{k}^{t})\pi_\theta(\mathbf{k}^t) \Big) \quad (\mathbf{x}^{\leq t} \text{ and } \mathbf{y}^{<t} \text{ are given}) && (17)\\
&\approx p(\mathbf{y}^t|\mathbf{x}^{\leq t}, \mathbf{y}^{<t}), && (18)
\end{aligned}
$$

where $q_\phi(\mathbf{k}^i)$ is an approximated posterior distribution and $p_\theta(\mathbf{k}^i)$ is a true posterior distribution.

B Training Details

All the parameters except pretrained parts are initialized with Xavier method (Glorot & Bengio, 2010). We use Adam optimizer (Kingma & Ba, 2015) with $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e-07$. For the models without BERT, we set the learning rate to 0.001 and initialize the embedding matrix with `fastText` (Bojanowski et al., 2016) trained on the Common Crawl corpus. For the models with BERT, we set the learning rate to 0.00002 and initialize encoder weights with `BERT-Base, Uncased` pretrained weights. We apply label smoothing (Pereyra et al., 2017; Edunov et al., 2017; Vaswani et al., 2017) for both knowledge selection and utterance generation, and set 0.1 and 0.05 for each. We set the temperature of Gumbel-Softmax to $\tau = 0.1$ and the hyperparameter for the knowledge loss to $\lambda = 1.0$. For efficiency, we batch the dialogues rather than individual turns. We train our model up to 5 epochs on a single NVIDIA TITAN Xp GPU.