Thinking Globally, Acting Locally: Distantly Supervised Global-to-Local Knowledge Selection for Background Based Conversation

Pengjie Ren[1], **Zhumin Chen**[2], **Christof Monz**[1], **Jun Ma**[2], **Maarten de Rijke**[1]
[1] University of Amsterdam, Amsterdam, The Netherlands
[2] Shandong University, Jinan, China
{p.ren, c.monz, derijke}@uva.nl, {chenzhumin, majun}@sdu.edu.cn

Abstract

Background Based Conversations (BBCs) have been introduced to help conversational systems avoid generating overly generic responses. In a BBC, the conversation is grounded in a knowledge source. A key challenge in BBCs is Knowledge Selection (KS): given a conversational context, try to find the appropriate background knowledge (a text fragment containing related facts or comments, etc.) based on which to generate the next response. Previous work addresses KS by employing attention and/or pointer mechanisms. These mechanisms use a *local* perspective, i.e., they select a token at a time based solely on the current decoding state. We argue for the adoption of a *global* perspective, i.e., pre-selecting some text fragments from the background knowledge that could help determine the topic of the next response. We enhance KS in BBCs by introducing a Global-to-Local Knowledge Selection (GLKS) mechanism. Given a conversational context and background knowledge, we first learn a topic transition vector to encode the most likely text fragments to be used in the next response, which is then used to guide the local KS at each decoding timestamp. In order to effectively learn the topic transition vector, we propose a distantly supervised learning schema. Experimental results show that the GLKS model significantly outperforms state-of-the-art methods in terms of both automatic and human evaluation. More importantly, GLKS achieves this without requiring any extra annotations, which demonstrates its high degree of scalability.

Introduction

Non-task-oriented conversational systems (a.k.a., chatbots) aim to engage users in conversations for entertainment (Yan 2018) or to provide valuable information (Zhou, Prabhumoye, and Black 2018). Sequence-to-sequence models are an effective framework that is commonly adopted in this field. However, a problem of vanilla sequence-to-sequence based methods is that they tend to generate generic and non-informative responses with bland and deficient responses (Chen et al. 2017). Various methods have been proposed to alleviate this issue, such as adjusting objective functions (Li et al. 2016; Zhang et al. 2018b; Jiang et al. 2019) or incorporating personal profiles (Zhang et al. 2018a).

Background Based Conversations (BBCs) have demonstrated a potential for generating more informative responses (Zhou, Prabhumoye, and Black 2018). Given some background knowledge (e.g., an article in the form of free text) and a conversation, the BBC task is to generate responses by referring to the background knowledge and considering the dialogue history context at the same time. A key challenge in BBCs is *Knowledge Selection* (KS), which is the task of finding the appropriate background knowledge (e.g., a text fragment about a movie plot) based on which the next response is to be generated.

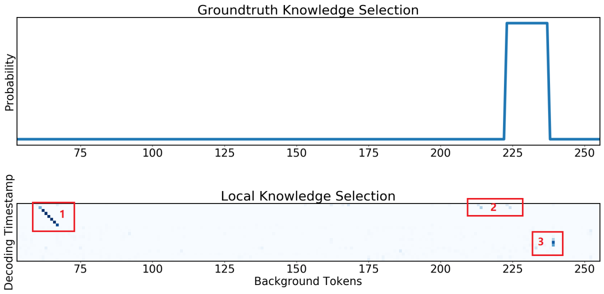


Figure 1: Visualization of local knowledge selection. The X-axes represent the background tokens; the top Y-axis represents KS probabilities and the spike indicates the ground truth KS; the bottom Y-axis represents the decoding timestamp and darker blue means larger KS probabilities.

Existing methods for BBCs can be grouped into two categories: extraction-based methods and generation-based methods. The former addresses KS by learning two pointers to extract spans from the background material as responses, and outperforms generation-based methods in finding knowledge (Moghe et al. 2018). However, there are two major issues with extraction-based methods. First, in most cases the generated responses are not natural due to their extractive nature. Second, unlike, e.g., Machine Reading Comprehension (MRC), in BBCs there is no notion of standard answer. For example, extraction-based methods cannot handle greetings in chitchats.

Today's generation-based methods perform KS with a local perspective, i.e., by selecting one token at a time based solely on the current decoding state. This is problematic because they lack the guidance that a more global perspective would offer. In Figure 1, we visualize the KS of a state-of-the-art model, an improved Get To The Point (GTTP), which achieves competitive performance on this task. The top figure corresponds to the ground truth KS annotations; the lower figure shows the KS probabilities of GTTP at each decoding timestamp. GTTP settles on two background areas (red box 1 and 2) at first in a sign of hesitation.

However, due to the lack of a global perspective, it chooses the wrong one (red box 1). And it is too late when GTTP realizes this and tries to correct its mistakes (red box 3). In this paper, we propose to address this issue and enhance KS for generation-based methods by introducing a *Global-to-Local Knowledge Selection* (GLKS) mechanism. The general idea is that we learn a "topic transition vector" with a Global Knowledge Selection (GKS) module beforehand, which sets the tone for the next response and encodes the general meaning of the most likely used background knowledge. The "topic transition vector" is then used to guide the Local Knowledge Selection (LKS) at each decoding timestamp to avoid situations like the one in Figure 1.

As in existing work, we train LKS with the Maximum Likelihood Estimation (MLE) loss. However, MLE is not effective enough to supervise the learning of GKS because it only provides token-wise supervision. To this end, we propose a distantly supervised learning schema where we use the Jaccard similarity between the ground truth responses and the background knowledge as an extra signal to train GKS. All parameters are learned by a linear combination of the global Distant Supervision (DS) and local MLE in an end-to-end back-propagation training paradigm.

Several recent studies try to improve the KS of generation-based methods. Meng et al. (2019) introduce a reference decoder that learns to directly select a semantic unit (e.g., a span containing complete semantic information) from the background, besides generating the response token by token. Liu et al. (2019) fuses two types of knowledge, triples from a structured knowledge graph and texts from unstructured background material, for better KS. Although they achieve promising improvements, they all have obvious limitations. Meng et al. (2019)'s work needs boundary annotations of semantic units in both backgrounds and responses to enable supervised training. To be able to put Liu et al. (2019)'s model to work, the authors prepare a structured knowledge source and manually ground unstructured background to it beforehand. To show the effectiveness of GLKS, we carry out experiments on the same datasets as Meng et al. (2019) and Liu et al. (2019). Our proposed GLKS model significantly outperforms their models as well as other state-of-the-art methods in terms of both automatic and human evaluation. GLKS is able to generate natural responses, yielding better KS, while requiring minimum efforts (in terms of human annotations), which means it exhibits better scalability.

Our contributions are summarized as follows:

- We propose a novel neural architecture with a Global-to-Local Knowledge Selection (GLKS) mechanism for BBCs that can generate more appropriate responses while retaining fluency.
- We devise an effective combined global (DS) and local (MLE) learning schema for GLKS without using extra annotations.
- Experiments show that GLKS outperforms state-of-the-art models by a large margin in terms of both automatic and human evaluation.

Related Work

Open-domain Conversation

Sequence-to-sequence modeling for open-domain conversations has been studied for years (Shang, Lu, and Li 2015). Previous studies have proposed various variants on different conversational tasks (Lowe et al. 2015; Serban et al. 2016) and have shown the superiority of sequence-to-sequence conversation modeling when compared to IR or template-based methods, especially in generating fluent responses. However, many challenges remain. Response informativeness is especially important; conversations become dull and less attractive due to too many generic responses such as "I don't know" and "I am sorry" (Vougiouklis, Hare, and Simperl 2016; He et al. 2017). A number of studies address this issue by promoting response diversity. They either propose new losses (Li et al. 2016; Zhao, Zhao, and Eskenazi 2017; Jiang et al. 2019) or introduce new learning schemas (Zhang et al. 2018b). Another strategy is to incorporate latent topic information (Xing et al. 2017) or leverage external knowledge (Ghazvininejad et al. 2018; Liu et al. 2018; Zhou et al. 2018; Young et al. 2018).

Background Based Conversation

Background Based Conversations (BBCs) have shown promising results in improving response informativeness (Zhou, Prabhumoye, and Black 2018; Dinan et al. 2019; Qin et al. 2019). Work on BBCs can be grouped into extraction-based and generation-based methods.

Extraction-based methods grew out of work on Machine Reading Comprehension (MRC), where a span is extracted from the background as response to a question (Seo et al. 2016). Extraction-based methods are good at locating the right background knowledge but because they are designed for MRC tasks, where user utterances are mostly simple questions that can be answered by a span, they are not suitable for BBCs (Moghe et al. 2018). The extracted spans are not natural as conversational responses, and in many cases there are no standard answers in BBCs, e.g., greeting chitchats or opinions.

Therefore, most recent studies on BBC focus on generation-based methods. Since generation-based methods can generate natural and fluent responses, the key challenge is to find the appropriate background knowledge (Lian et al. 2019). Zhang, Ren, and de Rijke (2019) introduce a preselection process that uses dynamic bi-directional attention to improve background KS by using the utterance history context as prior information. Li et al. (2019) devise an Incremental Transformer to encode multi-turn utterances along with background knowledge and design a two-pass decoder to improve KS. Meng et al. (2019) combine the advantages of extraction-based and generation-based methods by incorporating a reference decoder that learns to select a span from the background during decoding. Liu et al. (2019) combine two types of knowledge, triples from knowledge graphs and texts from unstructured documents. For KS, they use multi-hop walking on graphs, like Moon et al. (2019).

Unlike the work described above, we address KS in BBCs by introducing a novel Global-to-Local Knowledge Selection (GLKS) mechanism and a distantly supervised learning schema for better learning of the mechanism. Most importantly, the proposed GLKS needs neither span annotations

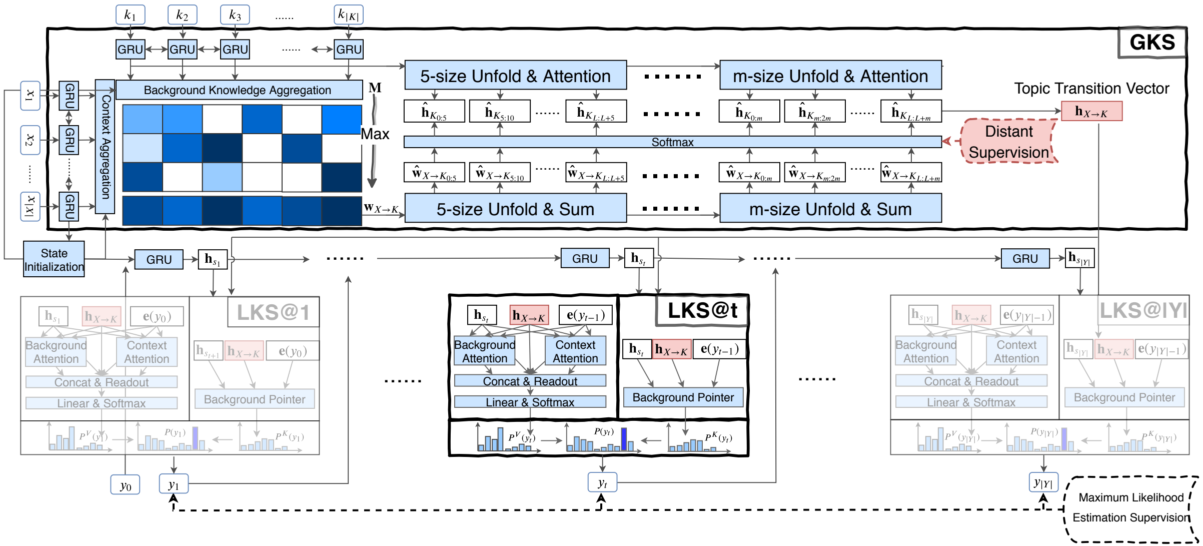


Figure 2: Overview of Global-to-Local Knowledge Selection (GLKS).

(Meng et al. 2019) nor extra knowledge grounding (Liu et al. 2019).

Global-to-Local Knowledge Selection

Given background material in the form of free text $K = [k_1, k_2, \ldots, k_t, \ldots, k_{|K|}]$, with $|K|$ tokens, and a current conversational context $X = [x_1, x_2, \ldots, x_t, \ldots, x_{|X|}]$, with $|X|$ tokens (usually, the previous n utterances), the task of BBC is to generate a response $Y = [y_1, y_2, \ldots, y_t, \ldots, y_{|Y|}]$ to X by occasionally referencing background knowledge in K.

The proposed model GLKS, shown in Figure 2, consists of four modules: *Background & Context Encoders*, a *Global Knowledge Selection (GKS) Module*, a *State Tracker*, and a *Local Knowledge Selection (LKS) Module*. Given K and X, the background & context encoders encode them into latent representations $\mathbf{H}^K$ and $\mathbf{H}^X$, respectively. Then, the GKS module evaluates the matching matrix between $\mathbf{H}^K$ and $\mathbf{H}^X$ globally. Based on the matching matrix, GKS makes a decision of "what to talk about next" by selecting continuous spans from the background K to form a "topic transition vector" $\mathbf{h}_{X\to K}$. At each decoding time step, LKS outputs a response token by either generating from the vocabulary or selecting from the background K under the guidance of the topic transition vector.

Background and context encoders

We use a bi-directional RNN with GRU (Cho et al. 2014) to convert the background and context sequences into two hidden representation sequences $\mathbf{H}^K = [\mathbf{h}^k_1, \mathbf{h}^k_2, \ldots, \mathbf{h}^k_t, \ldots, \mathbf{h}^k_{|K|}]$ and $\mathbf{H}^X = [\mathbf{h}^x_1, \mathbf{h}^x_2, \ldots, \mathbf{h}^x_t, \ldots, \mathbf{h}^x_{|X|}]$, respectively:

$$\mathbf{h}^k_t = \text{BiGRU}^K(\mathbf{e}(k_t), \mathbf{h}^k_{t-1}), \tag{1}$$

where $\mathbf{e}(k_t)$ is the token embedding vector; $\mathbf{h}^k_0$ is initialized with 0; $\mathbf{H}^X$ is obtained in a similar way but the BiGRU^X does not share parameters with BiGRU^K.

Global Knowledge Selection (GKS) module

Before calculating the match between $\mathbf{H}^K$ and $\mathbf{H}^X$, we first aggregate each representation in $\mathbf{H}^K$ and $\mathbf{H}^X$ with the last context output $\mathbf{h}^x_{|X|}$ using highway transformations (Srivastava, Greff, and Schmidhuber 2015):

$$\begin{aligned}\mathbf{h}^k_t &= g^k(\mathbf{W}_{linear}[\mathbf{h}^k_t, \mathbf{h}^x_{|X|}] + b) \\ &\quad + (1 - g^k)\tanh(\mathbf{W}_{non\text{-}linear}[\mathbf{h}^k_t, \mathbf{h}^x_{|X|}] + b), \\ g^k &= \sigma(\mathbf{W}_{gate}[\mathbf{h}^k_t, \mathbf{h}^x_{|X|}] + b]),\end{aligned} \tag{2}$$

where $\mathbf{W}_{linear}$, $\mathbf{W}_{non\text{-}linear}$ and $\mathbf{W}_{gate}$ are parameters; b is a bias; and σ is the sigmoid activation function. We formulate background knowledge aggregation as above. Context aggregation is achieved in a similar way. Both aggregations can be performed multiple times so as to get deep representations.

Next, we estimate the transition matching matrix $\mathbf{M} \in \mathbb{R}^{|K|\times|X|}$ between $\mathbf{H}^K$ and $\mathbf{H}^X$, each element of which is calculated as follows:

$$\mathbf{M}[i, j] = \mathbf{v}^{\mathrm{T}}_M \tanh(\mathbf{W}_{M1}\mathbf{h}^k_i + \mathbf{W}_{M2}\mathbf{h}^x_j), \tag{3}$$

where $\mathbf{v}_M$, $\mathbf{W}_{M1}$ and $\mathbf{W}_{M2}$ are parameters. We apply max pooling along the X dimension to get the transition weight vector $\mathbf{w}_{X\to K} \in \mathbb{R}^{|K|}$:

$$\mathbf{w}_{X\to K} = \max_X(\mathbf{M}). \tag{4}$$

Each element of $\mathbf{w}_{X\to K}$ represents the transition possibility w.r.t. the corresponding token in K.

The weight vector $\mathbf{w}_{X\to K}$ only considers token-wise transitions. However, a single token cannot determine the general meaning of the next response due to a lack of a global perspective. To address this, we introduce the "m-size unfold & sum" operation (as shown in Figure 2), which first extracts sliding adjacent weights of $\mathbf{w}_{X\to K}$ with an m-size window, and then sums them up. Specifically, each element of the semantic unit transition weight vector $\hat{\mathbf{w}}_{X\to K} =$

$[\hat{\mathbf{w}}_{X\to K_{0:m}}, \ldots, \hat{\mathbf{w}}_{X\to K_{L:L+m}}, \ldots]$ is calculated as follows:

$$\hat{\mathbf{w}}_{X\to K_{L:L+m}} = \sum_{i=L}^{L+m} \mathbf{w}_{X\to K}[i]. \tag{5}$$

We assume there is no overlap between two adjacent semantic units, which helps to reduce the size of $\hat{\mathbf{w}}_{X\to K}$.

Correspondingly, we fuse the "m-size unfold & attention" operation to obtain the semantic unit representations $\hat{\mathbf{H}}^K = [\hat{\mathbf{h}}_{K_{0:m}}, \ldots, \hat{\mathbf{h}}_{K_{L:L+m}}, \ldots]$ from $\mathbf{H}^K$:

$$\begin{aligned} \hat{\mathbf{h}}_{K_{L:L+m}} &= \sum_{i=L}^{L+m} \alpha_i \mathbf{h}_i^k \\ \alpha_i &= \text{attention}(\mathbf{h}^x_{|X|}, [\mathbf{h}^k_L, \ldots, \mathbf{h}^k_{L+m}]), \end{aligned} \tag{6}$$

where α_i is the additive attention weight between $\mathbf{h}^x_{|X|}$ and $\mathbf{h}^k_i$ (Bahdanau, Cho, and Bengio 2015). Note that α_i is normalized to probabilities with a local softmax operation (within the m-size window). Each $\hat{\mathbf{h}}_{K_{L:L+m}}$ corresponds to a semantic unit (a text fragment) $K_{L:L+m}$ in background K.

Finally, we get the topic transition vector $\mathbf{h}_{X\to K}$ with a soft weighted average over $\hat{\mathbf{H}}^K$:

$$\begin{aligned} \mathbf{h}_{X\to K} &= \sum_{L} P(K_{L:L+m} \mid X)\hat{\mathbf{h}}_{K_{L:L+m}} \\ P(K_{L:L+m} \mid X) &\propto \text{softmax}(\hat{\mathbf{w}}_{X\to K_{L:L+m}}). \end{aligned} \tag{7}$$

State tracker

The state tracker is responsible for initializing the decoding state at the start and updating it at each following time step. We get the initial decoding state $\mathbf{h}^s_0$ as follows:

$$\mathbf{h}^s_0 = \mathbf{W}_s[\mathbf{h}^x_{|X|}, \mathbf{h}_{X\to K}] + b, \tag{8}$$

where $\mathbf{W}_s$ is the parameter and s is the bias.

For updating, we employ another GRU that takes the generated token and decoding state of the previous time step as input and outputs the updated decoding state:

$$\mathbf{h}^s_t = \text{GRU}(\mathbf{e}(y_{t-1}), \mathbf{h}^s_{t-1}). \tag{9}$$

Here, y_0 is set to a special token "<BOS>," which indicates the start of decoding.

Local Knowledge Selection (LKS) module

At each decoding time step, we use the LKS module to predict each token one by one by either generating from vocabulary (with probability $P^V(y_t)$) or selecting from background K (with probability $P^K(y_t)$) under the guidance of the topic transition vector $\mathbf{h}_{X\to K}$, as shown in Figure 2.

Specifically, we first concatenate $\mathbf{h}_{X\to K}$, $\mathbf{h}^s_t$ and $\mathbf{e}(y_{t-1})$ to get the guidance vector $\mathbf{h}^g_t$ at t:

$$\mathbf{h}^g_t = [\mathbf{h}_{X\to K}, \mathbf{h}^s_t, \mathbf{e}(y_{t-1})]. \tag{10}$$

Then, we employ background attention to get the guidance-aware background representation $\hat{\mathbf{h}}^K_t$ in Eq. 11:

$$\begin{aligned} \hat{\mathbf{h}}^K_t &= \sum_{i=1}^{|K|} \alpha^K_i \mathbf{h}^k_i, \\ \alpha^K_i &= \text{attention}(\mathbf{h}^g_t, [\mathbf{h}^k_1, \ldots, \mathbf{h}^k_{|K|}]). \end{aligned} \tag{11}$$

In a similar way, we obtain the guidance-aware context representation $\hat{\mathbf{h}}^X_t$ with context attention.

We then construct the readout feature vector $\hat{\mathbf{h}}^r_t$ as follows:

$$\hat{\mathbf{h}}^r_t = \mathbf{W}_r[\mathbf{e}(y_{t-1}), \mathbf{h}^s_t, \mathbf{h}_{X\to K}, \hat{\mathbf{h}}^K_t, \hat{\mathbf{h}}^X_t], \tag{12}$$

where $\mathbf{W}_r$ are the parameter and b is the bias. The readout feature vector is then passed through a linear layer to estimate $P^V(y_t)$ with a softmax layer over the vocabulary:

$$P^V(y_t) = \text{softmax}(\mathbf{W}_V \hat{\mathbf{h}}^r_t), \tag{13}$$

where $\mathbf{W}_V \in \mathbb{R}^{|V|\times|F|}$ are the parameters, $|V|$ is the vocabulary size, and $|F|$ the hidden size of the readout feature vector $\hat{\mathbf{h}}^r_t$.

For $P^K(y_t)$, we employ another background attention as in Eq. 11 to learn a pointer α^P_i as the probability of selecting a background token k_i.

Finally, we combine $P^V(y_t)$ and $P^K(y_t)$ as follows:

$$\begin{aligned} P(y_t) &= gP^V(y_t) + (1-g)\sum_{y_t\in K} P^K(y_t) \\ g &= \sigma(\mathbf{W}\mathbf{h}^s_t + b), \end{aligned} \tag{14}$$

where g is a learnable soft gate to switch between $P^V(y_t)$ and $P^K(y_t)$.

Learning

To maximize the prediction probability of the target response given the context and background, we design three objectives, namely the Maximum Likelihood Estimation loss, the Distant Supervision loss, and the Maximum Causal Entropy loss.

The *Maximum Likelihood Estimation* (MLE) loss, which is commonly used, is defined as follows:

$$\mathcal{L}_{mle}(\theta) = -\frac{1}{M}\sum_{m=1}^{M}\sum_{t=1}^{|Y|}\log P(y_t), \tag{15}$$

where θ are all the parameters of our model, and M is the number of training samples.

The MLE loss only provides token-wise supervisions that lack a global perspective. To address this, we define the *Distant Supervision* (DS) loss to supervise the learning of GKS (see Figure 2) as follows:

$$\begin{aligned} \mathcal{L}_{ds}(\theta) &= \frac{1}{M}\sum_{m=1}^{M} D_{KL}(P(\hat{\mathbf{H}}^K)\|Q(\hat{\mathbf{H}}^K)) \\ P(\hat{\mathbf{H}}^K) &= \text{softmax}(\hat{\mathbf{w}}_{X\to K}) \\ Q(\hat{\mathbf{H}}^K) &= \text{softmax}(\text{Jaccard}(\hat{K}, Y)), \end{aligned} \tag{16}$$

where $\hat{\mathbf{w}}_{X\to K}$ is the semantic unit transition weight vector (Eq. 5) and $\hat{\mathbf{H}}^K$ are the semantic unit presentations (Eq. 6); Y is the ground truth response; $\hat{K} = [K_{0:m}, \ldots, K_{L:L+m}, \ldots]$ which is obtained with the same unfold operation as in Eq. 5 or 6. D_{KL} is the KL-divergence, which is commonly used to measure the distance between two probability distributions; $P(\hat{\mathbf{H}}^K)$ are the estimated probabilities of selecting the semantic units of $\hat{\mathbf{H}}^K$, which are obtained by using a softmax over the semantic unit transition weight vector; and, finally, $Q(\hat{\mathbf{H}}^K)$ are the distant ground truth supervisions, which are obtained by calculating the Jaccard similarity between each semantic unit $K_{L:L+m}$ and the ground truth response Y.

Because $Q(\hat{\mathbf{H}}^K)$ is distance based, we use the *Maximum Causal Entropy* (MCE) loss to alleviate the negative effects of the noise introduced by imprecise $Q(\hat{\mathbf{H}}^K)$:

$$\mathcal{L}_{mce}(\theta) = \frac{1}{M} \sum_{m=1}^{M} \sum_{t=0}^{|Y|} \sum_{w \in V} P(y_t = w) \log P(y_t = w). \quad (17)$$

The final loss is a linear combination of the three loss functions:

$$\mathcal{L}(\theta) = \mathcal{L}_{mle}(\theta) + \mathcal{L}_{ds}(\theta) + \mathcal{L}_{mce}(\theta). \quad (18)$$

All parameters of GLKS are learned in an end-to-end backpropagation training paradigm.

Experimental Setup

Implementation details

For a fair comparison, we stay close to previous studies regarding hyper-parameters. We set the word embedding size and hidden state size to 300 and 256, respectively. The word embeddings are initialized with GloVe (Liu et al. 2019). The vocabulary size is limited to ≈26,000. We limit the context length of all models to 65 (Moghe et al. 2018; Meng et al. 2019). We select the best models of all methods according to the validation set. We use gradient clipping with a maximum gradient norm of 2. We use the Adam optimizer ($\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$). We pre-train our model with the $\mathcal{L}_{ds}(\theta)$ loss for 10 epochs and then jointly train it with the other two losses. The code is available online.[1]

Dataset

We choose the Holl-E dataset released by Moghe et al. (2018) for experiments, because it is commonly used and contains the necessary information (boundary annotations, factoid knowledge) required by some recent methods (Meng et al. 2019; Liu et al. 2019). It contains ground truth KS labels that allow us to analyze the performance of models. Holl-E is built for movie chats in which each response is explicitly generated by copying and/or modifying sentences from the background (Moghe et al. 2018). The background consists of plots, comments and reviews about movies collected from different websites. Holl-E has three versions according to the background: oracle background (256-word), mixed-short background (256-word), and mixed-long background (1,200-word). Oracle background has just one kind of background information (plots, comments, etc.). We follow the original data split for training, validation and test, which contain 34,486, 4,388, and 4,318 samples respectively. There are two versions of the test set: one with a single golden reference (SR), the other with multiple golden references (MR).

Baseline

We compare with all generation-based methods for which results on the Holl-E dataset have been reported at the time of writing:

- **S2S** is a vanilla sequence-to-sequence model.
- **HRED** considers hierarchical modeling of context (Serban et al. 2016).
- **S2SA** fuses an attention mechanism to do KS at each decoding timestamp (Bahdanau, Cho, and Bengio 2015).
- **GTTP** leverages a copying/pointer mechanism together with an attention mechanism to do KS at each decoding timestamp (See, Liu, and Manning 2017).
- **Cake** introduces a pre-selection process that uses dynamic bi-directional attention to improve KS (Zhang, Ren, and de Rijke 2019).
- **RefNet** combines the advantages of BiDAF (Seo et al. 2016) and GTTP (See, Liu, and Manning 2017) by either selecting a span from the background with a reference decoder or generating a token with a generation decoder (Meng et al. 2019).
- **AKGCM** considers structured and unstructured knowledge for better KS (Liu et al. 2019). It uses policy network for KS on structured knowledge and GTTP for KS on unstructured knowledge and response generation.

S2S and HRED do not use any background knowledge; RefNet needs extra span annotations; AKGCM uses a structured knowledge graph and needs to manually ground knowledge between structured and unstructured sources.

Evaluation metrics

We use ROUGE-1, ROUGE-2 and ROUGE-L as automatic evaluation metrics.[2] Because the conversations are constrained by the background material, ROUGE scores are reliable. Nevertheless, we also randomly sample 500 test samples to conduct human evaluations on Amazon Mechanical Turk. For each sample, we show the responses from all systems to 3 workers and ask them to select all that are good[3] in terms of four aspects: (1) *Naturalness* (**N**), i.e., whether the responses are conversational, natural and fluent; (2) *Informativeness* (**I**), i.e., whether the responses use some background information; (3) *Appropriateness* (**A**), i.e., whether the responses are appropriate/relevant to the given context; and (4) *Humanness* (**H**), i.e., whether the responses look like they are written by a human.

Results

Automatic evaluation

The results of all methods on different settings (oracle, mixed-short and mixed-long) are shown in Table 1.

First, generally, GLKS achieves the best results on all metrics. GLKS significantly outperforms two recent best performing methods (RefNet and AKGCM) on the mixed-short background. The improvements show that GLKS is much better at leveraging and locating the right background information despite GLKS not using any extra annotations

[1] https://github.com/PengjieRen/GLKS

[2] We leave out BLEU since both previous and our experiments show that it has consistent performance with ROUGE (Moghe et al. 2018; Meng et al. 2019)

[3] We allow for an "all bad" option.

Table 1: Automatic evaluation results (%). **Bold face** indicates leading results in terms of the corresponding metric. Significant improvements over RefNet are marked with * (t-test, $p < 0.01$). SR and MR refer to test sets with single and multiple references. CaKe cannot run on the 1200-word background due to out of memory errors even with very small batch sizes (Zhang, Ren, and de Rijke 2019). The results of AKGCM are taken from the paper because the authors have not released their code.

	ROUGE-1		ROUGE-2		ROUGE-L	
	SR	MR	SR	MR	SR	MR
no background						
S2S	27.15	30.91	09.56	11.85	21.48	24.81
HRED	24.55	25.38	07.61	08.35	18.87	19.67
oracle background (256-word)						
S2SA	27.97	32.65	14.50	18.22	23.23	27.55
GTTP	29.82	35.08	17.33	22.00	25.08	30.06
CaKe	42.82	48.65	30.37	36.54	37.48	43.21
RefNet	42.87	49.64	30.73	38.15	37.11	43.77
GLKS	**43.75***	**50.67***	**31.54***	**39.20***	**38.69***	**45.64***
mixed-short background (256-word)						
S2SA	26.36	30.76	13.36	16.69	21.96	25.99
GTTP	30.77	36.06	18.72	23.70	25.67	30.69
CaKe	41.26	45.81	29.43	34.00	36.01	40.79
RefNet	41.33	47.00	31.08	36.50	36.17	41.72
AKGCM	–	–	29.29	–	34.72	–
GLKS	**44.52***	**50.06***	**33.05***	38.87*	**39.63***	**45.12***
mixed-long background (1,200-word)						
S2SA	21.90	24.90	5.63	7.00	17.02	19.65
GTTP	23.64	28.81	10.11	14.34	17.60	22.04
RefNet	34.90	42.08	**22.12**	**29.74**	29.64	36.65
GLKS	**35.30**	**42.31**	21.86	29.35	**30.36**	**37.30**

Table 2: Human evaluation results. $\geq n$ means that at least n MTurk workers think it is a good response w.r.t. *Naturalness* (**N**), *Informativeness* (**I**), *Appropriateness* (**A**) and *Humanness* (**H**).

	Improved GTTP		**RefNet**		**GLKS**	
	≥1	≥2	≥1	≥2	≥1	≥2
N	307	115	391	213	**424**	**226**
I	271	89	**411**	**244**	401	199
A	318	111	371	180	**406**	**219**
H	332	123	394	225	**436**	**263**

Table 3: Ablation study (%). -GKS, -$\mathcal{L}_{ds}(\theta)$ and -$\mathcal{L}_{mce}(\theta)$ denote GLKS without the corresponding part.

	ROUGE-1		ROUGE-2		ROUGE-L	
	SR	MR	SR	MR	SR	MR
-GKS	41.80	47.08	29.88	35.31	36.91	42.10
-$\mathcal{L}_{ds}(\theta)$	41.27	46.96	29.49	35.40	36.47	42.12
-$\mathcal{L}_{mce}(\theta)$	43.69	48.84	32.30	37.54	38.79	43.86
GLKS	**44.52**	**50.06**	**33.05**	**38.87**	**39.63**	**45.12**

(such as the span annotations used by RefNet) or information (such as the structured knowledge used by AKGCM). We analyze the improvements of GLKS in depth with an ablation study.

Second, the improvements of GLKS on the oracle and mixed-short background are much larger than on the mixed-long background. The reason is that KS becomes much more difficult when the background becomes longer. This is supported by the fact that the results of all methods drop around 10% compared with their results on the mixed-short background. This also means that there is still a long way to go for BBCs. GLKS and RefNet are comparable in the mixed-long background setting. GLKS only gains around 0.3% (ROUGE-1) and 0.7% (ROUGE-L) improvement over RefNet. RefNet is slightly better than GLKS on ROUGE-2. This is because RefNet uses extra span annotations, which shows great superiority in this setting.

Human evaluation

We conduct human evaluations to further compare GLKS and two strong baselines. The results are shown in Table 2. The improved GTTP is equivalent to LKS in this paper. Both GLKS and RefNet are better than GTTP on *Naturalness* because GTTP frequently generates responses with no topics or irrelevant topics, which makes it difficult for mturk workers to assess the fluency. RefNet gets the best votes on *Informativeness* which means it invokes background knowledge more frequently. This is consistent with its modeling schema, which encourages the model to refer to background during generation. However, this does not mean GLKS can always locate the appropriate background knowledge. GLKS achieves the best result on *Appropriateness*, which means it is indeed better at KS and can generate responses with more appropriate/relevant topics. Unsurprisingly, GLKS gets the most votes on *Humanness* because its responses are more natural and appropriate.

Analysis

Ablation study

To analyze where the improvements of GLKS come from, we conduct an ablation study as shown in Table 3. Generally, all three parts (the GKS module, the DS $\mathcal{L}_{ds}(\theta)$, and the MCE $\mathcal{L}_{mce}(\theta)$) are helpful because removing any of them will decrease the results consistently. GKS and $\mathcal{L}_{ds}(\theta)$ are much more effective because they yield around 3% improvements. This supports the motivations of our work which proposes to incorporate global perspective with distant supervision into KS. $\mathcal{L}_{mce}(\theta)$) is introduced to alleviate the negative effects of the noise introduced by imprecise distant supervisions. The results of -$\mathcal{L}_{mce}(\theta)$ in Table 3 demonstrate its usefulness. Even after removing all these modules, GLKS still outperforms vanilla GTTP. This is because we optimize the architecture with helpful tricks, e.g., using context states to aggregate background and context representations (like in Eq. 2), combining multiple representations to construct the readout feature vector (Eq. 12), etc.

Hyper-parameter analysis

There is a hyper-parameter m that controls the unfolding window size in Eq. 5 and 6. We plot the ROUGE scores on the validation and test sets in Fig. 3 to analyze its sensitivity. The ROUGE scores increase and decrease within the scope of around 2% difference which means GLKS is not sensitive

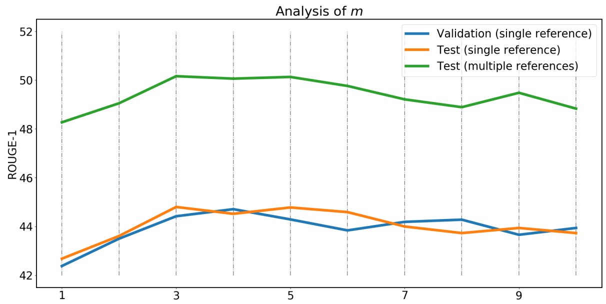
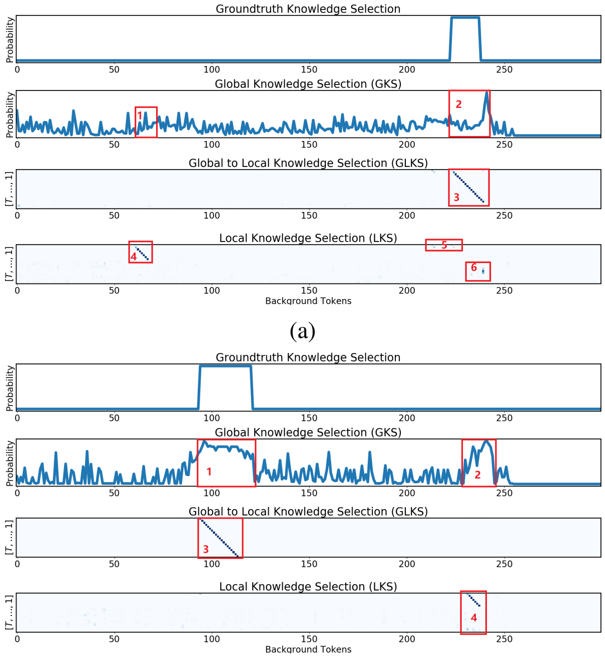


Figure 3: Analysis of m. The trends of ROUGE-2 and ROUGE-L are similar to ROUGE-1.



(a)

(b)

Figure 4: KS Visualization. For each figure, from top to bottom, are: ground truth KS, GKS, LKS and GLKS.

to m. The best results are achieved around $m = 3, 4, 5$ and the best validation results are achieved with $m = 4$. The results with $m \geq 3$ are much better than those with $m \leq 2$. Hence, m influences the performance and $m = 4$ is enough to discriminate different knowledge and guide KS.

Visual analysis

In Fig. 4 we visualize KS with different settings. The X axis corresponds to the background token sequence. The Y axis of the first and left two figures denote KS probabilities and decoding time steps, respectively. The color depth in the lower two figure represents token-wise KS probabilities.

We can see that without GKS, LKS can easily be fooled by similar but less appropriate knowledge (i.e., red box 1 and 2 in Figure 4a and 4b respectively). As a result, the model starts with the wrong or less appropriate knowledge (red box 4 in Figure 4a and 4b) or results in inconsistent KS (i.e., red box 4, 5, 6 in Figure 4a) during generation. This is because the model with only LKS lacks a global perspective as guidance, making it harder to make decisions and easier to make mistakes. In contrast, the model can avoid these issues and achieve better and more consistent KS when taking GKS into consideration (red box 3 in Fig. 4a and 4b).

Case study

We select an example from the test set to intuitively illustrate the responses generated by different models, as shown in Table 4. We can see that all models have learnt to invoke knowledge during generation. However, GTTP and LKS are relatively bad at KS, resulting in using less appropriate knowledge. RefNet is good at KS and can generate natural responses. But it has difficulties in coordinating the generation and reference decoding sometimes. As a result, it has a higher probability of generating contradictory responses. By comparison, GLKS can generate appropriate responses which yields better humanness.

Table 4: Case study.

	Backgound: ... later that evening , he intends to access kevin 's room , but kevin fools him into thinking that he has walked in on his father , causing the concierge to flee ... home alone 2 is a carbon copy , but it 's also much better and more complex a movie than the first ... regardless it 's a classic and i watch the first two movies every year ...
	H1: i loved all the tricks , and traps kevin created . **H2**: me too , i loved when using a tape recorder , he tapes a message and slows down his voice , placing a hotel reservation . **H1**: that was too funny , the hotel staff did n't believe him though .
GTTP	it 's a classic and i watch the first two movies every year .
RefNet	that it was so sad when he intends to access kevin 's room , but kevin fools him into thinking that he has walked in on his father , causing the concierge to flee .
LKS	i know , it was a carbon copy , but it 's also much better and more complex a movie than the first .
GLKS	so true , later that evening , he intends to access kevin 's room , but kevin fools him into thinking that he has walked in on his father , causing the concierge to flee .

There are also failure cases for GLKS as well as the other models: one severe issue is that the models tend to invoke the same knowledge even though the context has changed somewhat. This indicates that a mechanism is needed to track the already used knowledge.

Conclusion and Future Work

In this paper, we propose an end-to-end neural model for BBCs, which introduces a Global-to-Local Knowledge Selection (GLKS) mechanism to enhance KS. We also present a DS learning schema to learn GLKS effectively without using any extra annotations or information. Experiments show that with GLKS, our model can generate more appropriate and human-like responses.

As to future work, we intend to apply GLKS to other BBC tasks. Besides, GLKS can be advanced in many directions. First, better GKS modules can be designed to further improve KS especially when using very long background. Second, a mechanism can be incorporated into GLKS to enable the track of used knowledge in the context.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. This work is supported by Ahold Delhaize, the Association of Universities in the Netherlands (VSNU), the Innovation Center for Artificial Intelligence (ICAI), the Natural Science Foundation of China (61672324, 61672322, 61972234, 61902219), the Natural Science Foundation of Shandong province (2016ZRE27468), the Tencent AI Lab Rhino-Bird Focused Research Program (JR201932), and the Fundamental Research Funds of Shandong University. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Chen, H.; Liu, X.; Yin, D.; and Tang, J. 2017. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter* 19(2):25–35.

Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, 1724–1734.

Dinan, E.; Roller, S.; Shuster, K.; Fan, A.; Auli, M.; and Weston, J. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *ICLR*.

Ghazvininejad, M.; Brockett, C.; Chang, M.-W.; Dolan, B.; Gao, J.; Yih, W.-t.; and Galley, M. 2018. A knowledge-grounded neural conversation model. In *AAAI*, 5110–5117.

He, S.; Liu, C.; Liu, K.; and Zhao, J. 2017. Generating natural answers by incorporating copying and retrieving mechanisms in sequence-to-sequence learning. In *ACL*, 199–208.

Jiang, S.; Ren, P.; Monz, C.; and de Rijke, M. 2019. Improving neural response diversity with frequency-aware cross-entropy loss. In *The Web Conference*, 2879–2885.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL*, 110–119.

Li, Z.; Niu, C.; Meng, F.; Feng, Y.; Li, Q.; and Zhou, J. 2019. Incremental transformer with deliberation decoder for document grounded conversations. In *ACL*, 12–21.

Lian, R.; Xie, M.; Wang, F.; Peng, J.; and Wu, H. 2019. Learning to select knowledge for response generation in dialog systems. *arXiv*.

Liu, S.; Chen, H.; Ren, Z.; Feng, Y.; Liu, Q.; and Yin, D. 2018. Knowledge diffusion for neural dialogue generation. In *ACL*, 1489–1498.

Liu, Z.; Niu, Z.-Y.; Wu, H.; and Wang, H. 2019. Knowledge aware conversation generation with explainable reasoning on augmented graph. *arXiv*.

Lowe, R.; Pow, N.; Serban, I. V.; and Pineau, J. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *SIGDIAL*, 285–294.

Meng, C.; Ren, P.; Chen, Z.; Monz, C.; Ma, J.; and de Rijke, M. 2019. RefNet: A reference-aware network for background based conversation. *arXiv*.

Moghe, N.; Arora, S.; Banerjee, S.; and Khapra, M. M. 2018. Towards exploiting background knowledge for building conversation systems. In *EMNLP*, 2322–2332.

Moon, S.; Shah, P.; Kumar, A.; and Subba, R. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *ACL*, 845–854.

Qin, L.; Galley, M.; Brockett, C.; Liu, X.; Gao, X.; Dolan, B.; Choi, Y.; and Gao, J. 2019. Conversing by reading: Contentful neural conversation with on-demand machine reading. In *ACL*, 5427–5436.

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*, 1073–1083.

Seo, M.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2016. Bidirectional attention flow for machine comprehension. In *ICLR*.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A. C.; and Pineau, J. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, 3776–3784.

Shang, L.; Lu, Z.; and Li, H. 2015. Neural responding machine for short-text conversation. In *ACL*, 1577–1586.

Srivastava, R. K.; Greff, K.; and Schmidhuber, J. 2015. Training very deep networks. In *NeurIPS*, 2377–2385.

Vougiouklis, P.; Hare, J.; and Simperl, E. 2016. A neural network approach for knowledge-driven response generation. In *COLING*, 3370–3380.

Xing, C.; Wu, W.; Wu, Y.; Liu, J.; Huang, Y.; Zhou, M.; and Ma, W.-Y. 2017. Topic aware neural response generation. In *AAAI*, 3351–3357.

Yan, R. 2018. Chitty-chitty-chat bot: Deep learning for conversational AI. In *IJCAI*, 5520–5526.

Young, T.; Cambria, E.; Chaturvedi, I.; Zhou, H.; Biswas, S.; and Huang, M. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *AAAI*, 4970–4977.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? In *ACL*, 2204–2213.

Zhang, Y.; Galley, M.; Gao, J.; Gan, Z.; Li, X.; Brockett, C.; and Dolan, B. 2018b. Generating informative and diverse conversational responses via adversarial information maximization. In *NeurIPS*, 1810–1820.

Zhang, Y.; Ren, P.; and de Rijke, M. 2019. Improving background based conversation with context-aware knowledge pre-selection. In *SCAI*.

Zhao, T.; Zhao, R.; and Eskenazi, M. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL*, 654–664.

Zhou, H.; Young, T.; Huang, M.; Zhao, H.; Xu, J.; and Zhu, X. 2018. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, 4623–4629.

Zhou, K.; Prabhumoye, S.; and Black, A. W. 2018. A dataset for document grounded conversations. In *EMNLP*, 708–713.