UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Improving Background Based Conversation with Context-aware Knowledge Pre-selection

Zhang, Y.; Ren, P.; de Rijke, M.

Publication date
2019

Document Version
Final published version

Link to publication

Citation for published version (APA):
Zhang, Y., Ren, P., & de Rijke, M. (2019). *Improving Background Based Conversation with Context-aware Knowledge Pre-selection*. Paper presented at 4th International Workshop on Search-Oriented Conversational AI (SCAI), Macao, China. https://arxiv.org/abs/1906.06685

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:08 Oct 2026

Improving Background Based Conversation with Context-aware Knowledge Pre-selection

Yangjun Zhang , **Pengjie Ren** and **Maarten de Rijke**
University of Amsterdam
{y.zhang6, p.ren, derijke}@uva.nl

arXiv:1906.06685v1 [cs.CL] 16 Jun 2019

Abstract

Background Based Conversations (BBCs) have been developed to make dialogue systems generate more informative and natural responses by leveraging background knowledge. Existing methods for BBCs can be grouped into two categories: *extraction-based* methods and *generation-based* methods. The former extract spans from background material as responses that are not necessarily natural. The latter generate responses that are natural but not necessarily effective in leveraging background knowledge. In this paper, we focus on *generation-based* methods and propose a model, namely Context-aware Knowledge Pre-selection (CaKe), which introduces a pre-selection process that uses dynamic bi-directional attention to improve knowledge selection by using the utterance history context as prior information to select the most relevant background material. Experimental results show that our model is superior to current state-of-the-art baselines, indicating that it benefits from the pre-selection process, thus improving informativeness and fluency.

1 Introduction

Dialogue systems have attracted great attention recently. Many methods have achieved promising results [Sutskever *et al.*, 2014; Bahdanau *et al.*, 2015]. However, many challenges, including lack of diversity, limited informativeness and inconsistency, still remain. Among them, the lack of informativeness may well be the most important one. The responses of a chat bot may be diverse without being informative (e.g., *I can't tell you*, *I'm not sure*, *I don't have a clue*). Background Based Conversation (BBC) is a method that utilizes external knowledge to solve this problem, thus decreasing the number of bland and deflective responses [Moghe *et al.*, 2018]. Given background material and a conversation, the Background Based Conversation (BBC) task is to generate responses by referring to background information and considering the dialogue history context at the same time. The background material may be semi-structured information or free text; in this paper, background material is provided in the form of free text. BBCs have demonstrated a potential for generating more informative responses.

There are two main approaches to BBCs, extraction-based methods and generation-based methods. Extraction-based methods extract text segments from the background as responses. Hence, the responses are informative. Generation-based methods generate sequences based on a encoder-decoder mechanism and the responses are natural and fluent. For extraction-based methods, models including the Bi-directional attention flow (BiDAF) model [Seo *et al.*, 2017] have been proposed for selecting the best matching positions of the tokens from the target context. However, responses produced by extraction-based methods are directly copied from background sentences. As a result, they are neither fluent nor natural. Generation-based methods include Hierarchical Recurrent Encoder-decoder Architecture (HRED) [Serban *et al.*, 2016] and Get To The Point (GTTP) [See *et al.*, 2017; Moghe *et al.*, 2018]; they are not always effective at leveraging background knowledge and may return responses with inappropriate background knowledge.

The motivation behind our model is to improve the process of utilizing background knowledge. The GTTP model [See *et al.*, 2017; Moghe *et al.*, 2018] selects background knowledge by using a hidden state at each decoding time step as a query to select background knowledge. For each query, it is supposed to comprise features of the corresponding response token, features from the previous response tokens, as well as features passed on from the utterance history context. However, the query may not contain all information from the utterance history since the information attenuated rapidly [Cho *et al.*, 2014] during state transfer especially when the length of the input utterance history context of the encoder increases, as the red line demonstrates in Figure 1.

The utterance history contains rich information about the conversation and it is also highly relevant to the response. Therefore, we propose a neural model, namely *Context-aware Knowledge Pre-selection* (CaKe), which introduces a knowledge pre-selection step to offset the problem of the current state-of-art model. CaKe utilizes utterance history context as query directly to facilitate better selection of knowledge. There are two steps to implement knowledge pre-selection in CaKe. To start, we choose the encoder state of the utterance history context as a query to select the most relevant knowledge. Then, we employ a modified version of BiDAF

Proceedings of the 2019 IJCAI Workshop SCAI: The 4th International Workshop on Search-Oriented Conversational AI

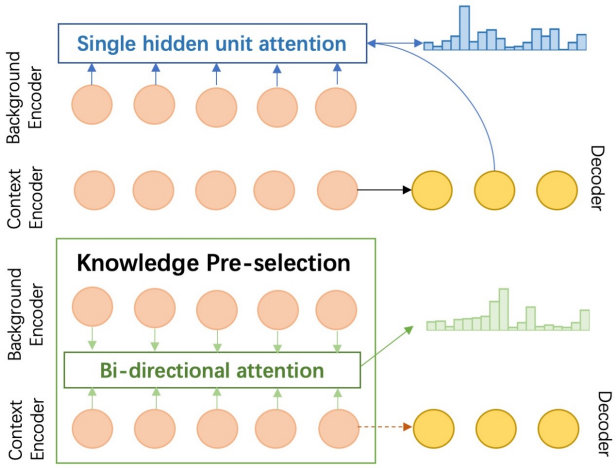

Figure 1: Knowledge pre-selection.

that combines background-to-context attention and context-to-background attention to point out the most relevant token positions of the background sequence to cope with long background text.

The main contribution of our model is to propose a pre-selection module. The pre-selection process leverages the utterance history of the corresponding response as a query to search a set of positions of the background knowledge sentence and find the most relevant knowledge. Meanwhile, we use a modified bi-attention flow mechanism to achieve the pre-selection process. Experiments show that CaKe significantly outperforms the strong generation-based baseline, GTTP. Although CaKe cannot beat extraction-based baseline, BiDAF, to a large margin, CaKe can generate more natural responses than BiDAF.

2 Related Work

2.1 Extraction-based Methods

Extraction-based methods to BBC are originally derived from Reading Comprehension (RC) tasks [Rajpurkar *et al.*, 2016], where the answer could be picked from a set of token positions of the input sequence. Vinyals *et al.* [2015] propose a Pointer Network (Ptr-Net) model that uses attention as a pointer to select a token in an input sequence so as to generate an output sequence where some tokens come from the input sequence. Wang and Jiang [2017] extend this work by combining match-Long Short-Term Memory (LSTM) and Ptr-Net. Seo *et al.* [2017] introduce BiDAF to improve the extraction of the context span by pointing to the start point and the end point of the relevant span. Lee *et al.* [2016] and Yu *et al.* [2017] predict answers by ranking text spans, where the token positions of the span are continuous, within background passages. Wang *et al.* [2017] predict the boundary of the answer span by a self-matching mechanism. Extraction-based methods are better at locating the right background span than generation-based methods [Moghe *et al.*, 2018]. Nevertheless, extraction-based methods are not suitable for BBCs as BBCs do not have standard answers like those in RC tasks and responses based on fixed extraction of the background are not natural or fluent enough for conversation tasks.

2.2 Generation-based Methods

Sequence to Sequence (S2S) learning methods [Sutskever *et al.*, 2014], later extended to Sequence to Sequence with Attention (S2SA) [Bahdanau *et al.*, 2015], are the basis of most generation-based methods. Although generation-based methods achieve good results on different conversation tasks, many challenges remain.

First, response diversity is to be improved. Lots of methods have been proposed to solve this issue. To increasediversity, Li *et al.* [2016] present Maximum Mutual Information (MMI) as the objective function in neural models to decrease generic response sequences and increase varied and interesting outputs. Serban *et al.* [2016] propose HRED, which uses a two-level hierarchy, including word level and dialogue turn level, to exploit long-term text. Later, Serban *et al.* [2017] add a high-dimensional stochastic latent variable to the target to extend HRED, aiming to generate responses with more diversified content and reduce blandness. Zhang *et al.* [2018b] explicitly optimize a variational lower bound on pairwise mutual information between query and response to boost diversity during training.

Second, response informativeness is an important issue. Many methods adding background knowledge and common sense to conversations have been proposed for mitigating blandness. Most of the existing conversational datasets are not labelled with relevant knowledge, so it is difficult to apply large datasets to the model training. As a result, most models need to do knowledge selection before training the model, such as knowledge diffusion [Liu *et al.*, 2018] and graph attention [Zhou *et al.*, 2018]. These methods use knowledge datasets that are separate from the conversations. Recently, several datasets have become available where conversations are generated based on background knowledge. Moghe *et al.* [2018] build a dataset for BBCs and conduct experiments with several methods. The datasets of Persona-chat [Zhang *et al.*, 2018a] and Wizard-of-Wikipedia [Dinan *et al.*, 2019] are similar to Moghe *et al.* [2018].

Although the background knowledge available in BBCs has a low degree of redundancy, selecting the most relevant background is important to improve informativeness. Moghe *et al.* [2018] employ GTTP proposed by See *et al.* [2017] to copy tokens from background knowledge at each generation timestamp. Lian *et al.* [2019] use posterior knowledge distribution to guide knowledge selection. The selection of background material plays a vital role in generating informative responses. However, the crucial role of context history in selecting appropriate background has not been fully explored by current methods. Unlike previously proposed methods, in order to encourage informative and non-deflective responses, our proposed model leverages the context as prior context to do pre-selection of the background knowledge.

3 Methodology

Given a background in the form of free text and the current utterance history context, BBC aims to generate an utterance as the next response. Formally, let $b = (b_1, b_2, \ldots, b_i \ldots, b_I)$ represent the words in the background knowledge, a current utterance history context in the form of

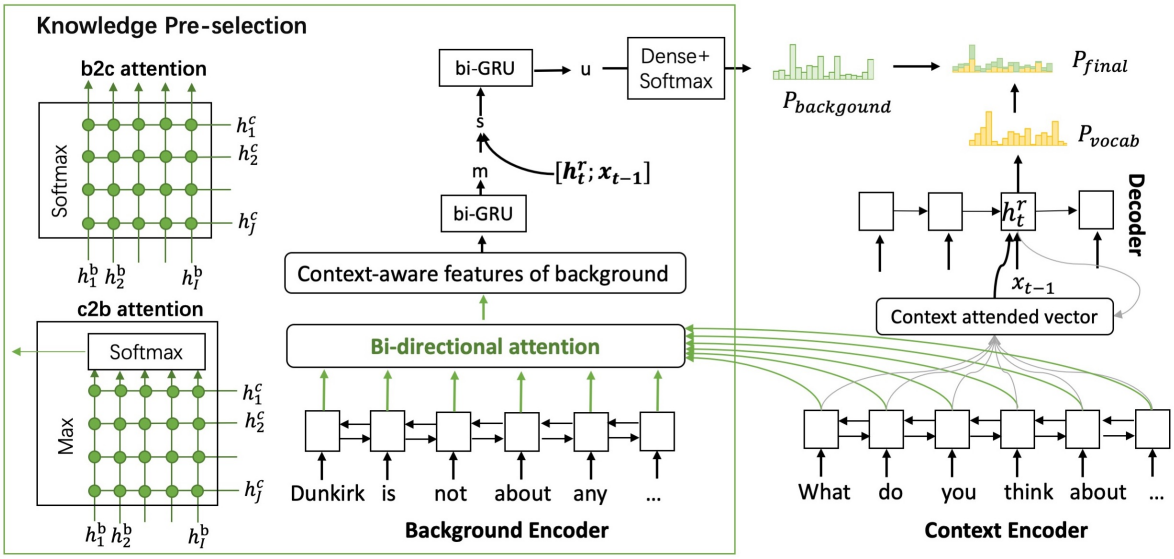

Figure 2: Overview of our model, Context-aware Knowledge Pre-selection (CaKe).

$c = (c_1, c_2, \ldots, c_j, \ldots, c_J)$, and the task of BBC is to generate response $r = (x_1, x_2, \ldots, x_t, \ldots, x_T)$ based on b and c.

In this section we introduce our model CaKe for BBC. The overview of CaKe is shown in Figure 2. First, we use two encoders to encode background and context. Second, for knowledge pre-selection, we select the background related to the context. To achieve this, we choose the encoder state of the context as query to select the most relevant knowledge from the background knowledge. We use a modified version of BiDAF [Seo *et al.*, 2017] which combines background-to-context attention and context-to-background attention to point out the most relevant token position of the background sequence. The pre-selection module forms context-aware background distribution. Third, for the generation part, the generator generates vocabulary distribution with global attention [Bahdanau *et al.*, 2015]. Lastly, We combine this pre-selector part with the generator to generate the final output of each decoding time step. We get the response based on the probability of generating token from the vocabulary or copying token from the background.

3.1 Background and Context Encoder

The word embeddings are used to map words to high-dimensional vector space. We apply random initialized word embeddings.

The background and context encoders encode background and context embeddings into $h^b = (h_1^b, h_2^b, \ldots, h_i^b, \ldots, h_I^b)$ and $h^c = (h_1^c, h_2^c, \ldots, h_j^c, \ldots, h_J^c)$ respectively. We use bidirectional Recurrent Neural Networks (RNNs) and concatenate the outputs of two RNNs for the two encoders. Therefore, we get $h^b \in \mathbb{R}^{2d \times I}$ for the background and $h^c \in \mathbb{R}^{2d \times J}$ for the context.

3.2 Knowledge Pre-selection

The selector is used for pre-selecting background words and form context-aware background distribution. We combine background-to-context and context-to-background attention. The original BiDAF [Seo *et al.*, 2017] trains start and end span position. Moghe *et al.* [2018] uses this model to find the model relevant span from the background knowledge by two positions, while we only use the start span position as the most relevant token index. The similarity score between context and background is calculated by:

$$score_{ij} = S(h_{:i}^b, h_{:j}^c), \tag{1}$$

where $h_{:i}^b$ is the i^{th} column vector of h^b and $h_{:j}^c$ is the j^{th} column vector of h^c. Meanwhile, S is defined as:

$$S(h^b, h^c) = w^T [h^b; h^c; h^b \odot h^c], \tag{2}$$

where w^T is a trainable weight vector. Then, we feed $score_{ij}$ to softmax function to get the attention and the corresponding vector.

First, background-to-context (b2c) attention reflects which context words are most relevant to each background word. α_i represents the attention weights on the context words by the i-th background word, where $\sum_i a_{ij} = 1$. The attention vector of context is computed by

$$\widetilde{h}_{:i}^c = \sum_j \alpha_{ij} h_{:j}^c, \tag{3}$$

where α_i is computed by normalizing $S_{i:}$ by a $\mathrm{softmax}$ function. $\widetilde{h}^c \in \mathbb{R}^{2d \times I}$.

Second, context-to-background (c2b) attention computes which background words are most relevant to each context word. The attention weights on the background words are calculated by $\beta = \mathrm{softmax}(\max_{col}(S)$. Then the attended background vector is computed by:

$$\widetilde{h}_{:i}^b = \sum_i \beta_i h_{:i}^b, \tag{4}$$

where $\widetilde{h}^b \in \mathbb{R}^{2d \times I}$.

Finally, contextual embeddings and the attention vectors are combined to yield g, where g is defined by:

$$g_{:i} = \eta(h^b_{:i}, \widetilde{h}^c_{:i}, \widetilde{h}^b_{:i}), \tag{5}$$

where η is a trainable vector. We use simple concatenation in our experiments: $\eta(h^b, h^c, \widetilde{h}^b) = [h^b, \widetilde{h}^c, h \odot \widetilde{h}^c, h^b \odot \widetilde{h}^b]$. g is the static context-aware background representations. We use a bi-RNN layer for g, and get m, which captures the interaction among the background words conditioned on the context. Then we concatenate m with h^r_t and x_{t-1} to generate s. s is fed into bi-RNN to generate u. Finally, background distribution is calculated by:

$$P_{background} = \text{softmax}(w^T_{p1}[g; m; s; u]), \tag{6}$$

where W^T_{p1} is a trainable factor.

3.3 Generator

For the generator, we generate the response token, which is the vocabulary distribution, with attention. For the generator module, the hidden state of response is: $h^r = (h^r_1, h^r_2, \ldots, h^r_t, \ldots, h^r_T)$, where h^r_t is the state of the decoder at the current time step. The final decoder hidden state aware representation of the context is the attention weighted sum of context c_t. The representation of context is $h^c_1, h^c_2, \ldots, h^c_j, \ldots, h^c_J$, where J is the total length of the context. The final attention weighted sum of the context is calculated as follows:

$$\begin{aligned} e^t_i &= v^T \tanh(W_c h^c_j + V h^r_t + b_c) \\ \gamma^t &= \text{softmax}(e^t) \\ c_t &= \sum_j \gamma^t_j h^c_j, \end{aligned} \tag{7}$$

where h^r_t is the current state of the decoder.

The generator then uses c_t, s_t and x_t to generate P_{vocab}, and the generation probability p_{gen} is calculated as follows:

$$P_{gen} = \sigma(w^T_c c_t + w^T_s s_t + w^T_x x_t + b_{gen}), \tag{8}$$

where $w^T_c, w^T_s, w^T_x, b_{gen}$ are trainable parameters. p_{gen} is used as a switch to choose between generating a word by sampling from the vocabulary distribution P_{vocab} or copying a word from the background by sampling from background distribution $P_{background}$.

3.4 Mixture and Loss

The mixture is used to mix the results of knowledge preselection and generation. We obtain the following final distribution over the extended vocabulary:

$$P_{final}(w) = p_{gen} P_{vocab}(w) + (1 - p_{gen}) P_{background}. \tag{9}$$

Let $P(w) = P_{final}(w)$. During training, the training loss of time step t is the negative log likelihood of the target word w^*_t for that time step:

$$loss_t = -\log P(w^*_t). \tag{10}$$

The training loss of the whole sequence $loss$ and the whole datasets $L(\theta)$ are:

$$\begin{aligned} loss &= \frac{1}{T} \sum_{t=0}^{T} loss_t \\ L(\theta) &= \sum_{n=0}^{N} loss, \end{aligned} \tag{11}$$

where θ is the set of all trainable weights, N is the number of total samples in the dataset.

4 Experiment Setup

4.1 Datasets

Recently, several datasets have been released for BBCs [Zhang *et al.*, 2018a; Moghe *et al.*, 2018; Dinan *et al.*, 2019]. We choose the Holl-E dataset released by Moghe *et al.* [2018]. The data contains background documents of 921 movies and 9071 conversations. The background documents of the movies contain four parts: review, plot, comment, meta-data or fact table. The conversations have two speakers. For the first speaker, the background documents are not available while the second one could use knowledge from background documents during chatting. We use two versions of background documents: oracle background and 256-word background. Oracle background uses the actual resource part from the background documents. The 256 words background is generated by truncating the background sentences.

4.2 Implementation Details

We use Gated Recurrent Units (GRU) [Chung *et al.*, 2014] as the RNN cell. The dimension of word embedding is 128 according to a rule of thumb and the GRU cell has a 256-dimensional hidden size. We use a vocabulary of 45k words. We limit the context length of all models to 120. We train all the models for 30 epochs and the loss converges. The best model is selected based on the BLEU and ROUGE metric. We use gradient clippings with a maximum gradient norm of 2 and do not use any form of regularization. The word embeddings are learned from scratch during training. We use Adam optimizer with batch size 32 and learning rate 0.001. CaKe is written in PyTorch and trained on four GeForce GTX TitanX GPUs.

4.3 Baselines and Evaluation Metrics

We compare CaKe with several state-of-the-art methods. **S2S** [Sutskever *et al.*, 2014] generates response from context with a simple sequence-to-sequence structure. **HRED** [Serban *et al.*, 2016] is a model using two-level hierarchy to encode the context. Seq2seq and HRED do not use background knowledge. **S2SA** is a model using attention mechanism to attend to background knowledge [Bahdanau *et al.*, 2015]. **GTTP** leverages background information with a copying mechanism to copy a token from the background at the appropriate decoding step [See *et al.*, 2017]. **BiDAF** extracts a span from the background as response and uses a co-attention architecture to improve the span finding accuracy [Seo *et al.*, 2017].

We use BLEU, ROUGE-1, ROUGE-2 and ROUGE-L as the automatic evaluation metrics. Because the background knowledge and the corresponding conversations are restricted to a specific topic, there automatic evaluations are relatively more reliable for BBCs than for open-domain conversational modeling [Dinan *et al.*, 2019].

5 Results and Analysis

5.1 Overall Performance

We list the result of all methods in Table 1.

Table 1: Automatic evaluation results.

Methods	BLEU		ROUGE-1		ROUGE-2		ROUGE-l	
	SR	MR	SR	MR	SR	MR	SR	MR
No background								
S2S	4.63	7.01	26.91	30.50	9.34	11.36	21.58	24.99
HRED	5.23	5.38	24.55	25.38	7.66	8.35	18.87	19.67
256 words background								
S2SA	11.71	12.76	26.36	30.76	13.36	16.69	21.96	25.99
BiDAF	**27.44**	**33.40**	38.79	43.92	**32.91**	**37.86**	35.09	40.12
GTTP	13.97	18.63	29.82	35.02	17.98	22.54	25.14	33.01
CaKe	26.17	29.49	**41.26**	**45.81**	29.43	34.00	**36.01**	**40.79**
Oracle background								
S2SA	12.26	13.11	27.51	31.89	13.98	17.55	22.85	27.03
BiDAF	24.93	**32.21**	35.60	42.40	29.48	**36.54**	31.72	38.39
GTTP	15.32	17.32	30.60	35.78	17.18	21.89	24.99	29.77
CaKe	**26.02**	31.16	**42.82**	**48.65**	**30.37**	**36.54**	**37.48**	**43.21**

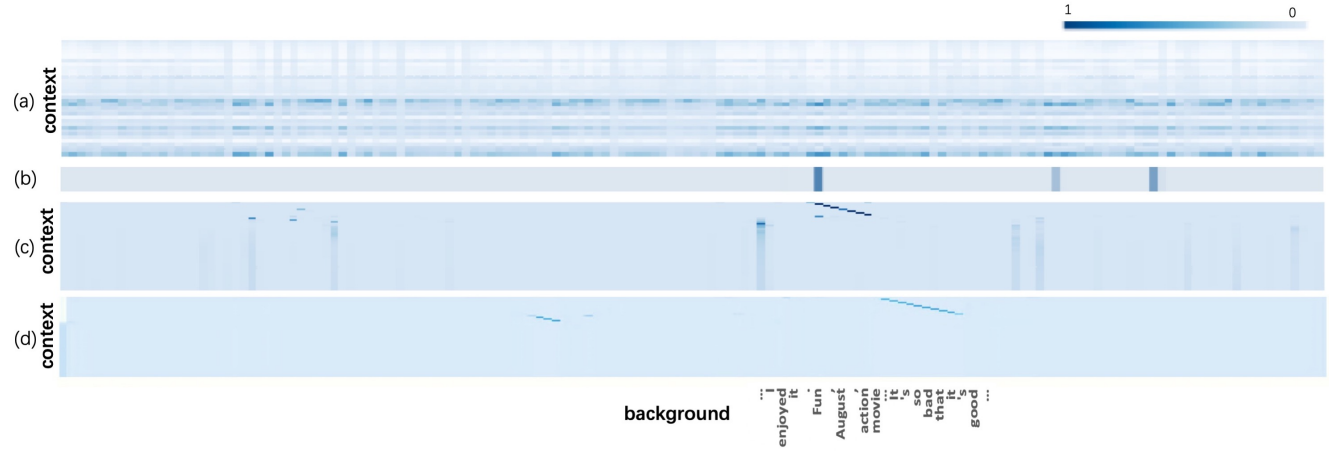

Figure 3: Knowledge selection visualization. (a) Background-to-context attention of CaKe; (b) Context-to-background attention of CaKe; (c) final pre-selection distribution on the background of CaKe; (d) knowledge distribution of GTTP.

First, CaKe outperforms the generation-based models, including S2S, HRED, S2SA and GTTP. Especially, for the strong baseline model GTTP, CaKe outperforms it by more than 35% for all the metrics with regard to the oracle background. Meanwhile, CaKe outperforms it by more than 20% for each metric when it comes to the 256 words background. The improvements show that CaKe is much better at locating the most relevant information in the background. The original GTTP model uses the current decoder state to select knowledge, while CaKe uses b2c and c2b attention to do pre-selection of knowledge. The comparison of CaKe and GTTP suggests that knowledge pre-selection is superior to classic single-state attention which uses the current state to attend to knowledge from the background.

Second, CaKe is superior to the BiDAF model with oracle background for all the metrics for more than 8% except BLEU and ROUGE-2 with multi-references. For the 256 words background, our model beats BiDAF on ROUGE-1 and ROUGE-l metrics. This suggests that CaKe can generate better responses than BiDAF. The main reason is that besides extracting knowledge from background like BiDAF, CaKe can also generate tokens from vocabulary to enhance response fluency.

Third, the performance of CaKe reduces slightly when the background becomes longer, but the reduction is acceptable considering that for the 256 words background CaKe is still slightly superior to BiDAF.

5.2 Knowledge Selection Visualization

As shown in Figure 3, we visualize the attention weights to highlight the differences between the results of GTTP and CaKe. The responses of GTTP and CaKe are: *'It was so bad that it's good'* and *'I agree, Fun, August, action movie'* respectively. The result of CaKe is closest to the ground truth.

The results of context-to-background attention show that attention is very strong on several positions of the background including the 92-th position, which is the token 'fun' and suggests that the utterance history could help find the most rel-

Table 2: Case studies.

Background	The Mist, what? A bit like The Fog, then. Stephen King's The Mist, oh, that makes it even worse. Directed by Frank Darabont, since when did he direct horror films? Okay, so he scripted Nightmare on Elm Street 3 and The Blob, not bad films, but not classics in any sense. Starring Thomas Jane, has anyone seen The Punisher. And, to cap it all, The Mist died a time. Love this movie, ooof that ending. Sometimes I feel like the only person who prefers the book ending. It's more expansive and leaves something to the imagination. Classic Horror in a Post Modern age. The ending was one of the best I've seen. 'The Mist' is worth watching! My favorite character was Melissa Mcbride. My favorite character was the main protagonist, David Drayton.
Context	Speaker 1: Which is your favorite character in this? Speaker 2: My favorite character was the main protagonist, David Drayton. Speaker 1: What about that ending?
Response	BiDAF: Classic horror in a post modern age. GTTP: They this how the mob mentality and religion turn people into monsters. CaKe: One of the best horror films I've seen in a long, long time.
Background	In many sites, even in IMDb, there are some pops-up screens that irritate me. What I do not understand is some comments of persons that should never watch this type of movie. What do they expect to see when they go to the movie theaters or buy/rent a DVD of 'Scary Movie 3', directed by David Zucker and with Leslie Nielsen in the cast? An art movie, with hidden messages, an epic, a classic or a film with politically correct jokes? Honestly, if I did not like this genre, in which Michael Jackson is disguised as an alien. My favorite character was Brenda. I think it was hilarious. Do you remember who came out of box office $110,000,082 awards BMI Film & TV Awards 2004 James L. Venable MTV Movie + TV Awards 2004 Best Cameo
Context	Speaker 1: And this again proved brilliance. Speaker 2: I totally loved this one. Tho the ending kinda weird but overall it gives me a creep. Speaker 1: Do you have any idea, how much it made on box office?
Response	BiDAF: $110,000,082. GTTP: It made $110,000,082. CaKe: I think it grossed $110,000,082.

evant positions of the background. It suggests that our pre-selection mechanism could help knowledge selection. This explains why our model could select relevant knowledge from the background better than GTTP.

5.3 Case Study

We select two examples from the test dataset to illustrate our results, as shown in Table 2. The examples suggest that CaKe is able to generate more fluent responses than BiDAF and more informative responses than GTTP. First, CaKe generates responses that are more relevant than GTTP. Second, the responses of CaKe are more natural and fluent than BiDAF, which can be inferred from the responses of BiDAF including *'Classic horror in a post modern age'* and *'$110,000,082'*. The reason is that BiDAF extracts spans from the background sentences as responses directly.

There are also occasions that CaKe does not perform well. For instance, CaKe generates common tokens like *'I agree'* and *'I know'* very frequently. This suggests that the diversity of the model needs to be taking into consideration.

6 Conclusion and Future Work

In this paper, we propose knowledge pre-selection process for the BBC task. The proposed model, CaKe, explores selecting relevant knowledge by using context as prior query. Experiments show that CaKe outperforms the state-of-art method.

A limitation of CaKe is that the performance of our pre-selection process decreases when the background becomes longer. To further improve CaKe in knowledge selection, we will explore alternative approaches to improve the selector and generator module in future work, such as multi-agent learning, transformer models and other attention mechanisms. As another future work, although automatic evaluations are relatively reliable for BBCs, it would be better if we can also conduct human evaluations. Meanwhile, we also hope to improve the diversity of CaKe by incorporating mechanisms such as changing optimization objects and leveraging mutual information [Jiang *et al.*, 2019].

Acknowledgments. This research was partially supported by Ahold Delhaize, the Association of Universities in the Netherlands (VSNU), the China Scholarship Council (CSC), and the Innovation Center for Artificial Intelligence (ICAI). All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

Code

To facilitate reproducibility of the results in this paper, we are sharing the code at https://github.com/repozhang/bbc-pre-selection.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015.

Kyunghyun Cho, Bart Van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, 2014.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*, 2019.

Shaojie Jiang, Pengjie Ren, Christof Monz, and Maarten de Rijke. Improving neural response diversity with frequency-aware cross-entropy loss. In *The Web Conference 2019*, pages 2879–2885, 2019.

Kenton Lee, Shimi Salant, Tom Kwiatkowski, Ankur Parikh, Dipanjan Das, and Jonathan Berant. Learning recurrent span representations for extractive question answering. *arXiv preprint arXiv:1611.01436*, 2016.

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

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

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. Knowledge diffusion for neural dialogue generation. In *The 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1498, 2018.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M Khapra. Towards exploiting background knowledge for building conversation systems. In *2018 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2332, 2018.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, 2016.

Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. In *The 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, 2017.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations*, 2017.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, 2014.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2692–2700, 2015.

Shuohang Wang and Jing Jiang. Machine comprehension using match-lstm and answer pointer. In *International Conference on Learning Representations*, 2017.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. Gated self-matching networks for reading comprehension and question answering. In *The 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198, 2017.

Yang Yu, Wei Zhang, Kazi Hasan, Mo Yu, Bing Xiang, and Bowen Zhou. End-to-end answer chunk extraction and ranking for reading comprehension. In *Thirty-First AAAI Conferenceon Artificial Intelligence*, 2017.

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 *The 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, 2018.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems*, pages 1810–1820, 2018.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. Commonsense knowledge aware conversation generation with graph attention. In *International Joint Conference on Artificial Intelligence*, pages 4623–4629, 2018.