RefNet: A Reference-Aware Network for Background Based Conversation

Meng, C.; Ren, P.; Chen, Z.; Monz, C.; Ma, J.; de Rijke, M.

DOI
10.1609/aaai.v34i05.6370

Publication date
2020

Document Version
Author accepted manuscript

Published in
AAAI-20, IAAI-20, EAAI-20 proceedings

Citation for published version (APA):

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, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
RefNet: A Reference-aware Network for Background Based Conversation

Chuan Meng,1 Pengjie Ren,2* Zhumin Chen,1* Christof Monz,2 Jun Ma,1 Maarten de Rijke2
1Shandong University, Qingdao, China, 2University of Amsterdam, Amsterdam, The Netherlands
mengchuan@mail.sdu.edu.cn
{p.ren, c.monz, derijke}@uva.nl, {chenzhumin, majun}@sdu.edu.cn

Abstract
Existing conversational systems tend to generate generic responses. Recently, Background Based Conversations (BBCs) have been introduced to address this issue. Here, the generated responses are grounded in some background information. The proposed methods for BBCs are able to generate more informative responses, however, they either cannot generate natural responses or have difficulties in locating the right background information. In this paper, we propose a Reference-aware Network (RefNet) to address both issues. Unlike existing methods that generate responses token by token, RefNet incorporates a novel reference decoder that provides an alternative way to learn to directly select a semantic unit (e.g., a span containing complete semantic information) from the background. Experimental results show that RefNet significantly outperforms state-of-the-art methods in terms of both automatic and human evaluations, indicating that RefNet can generate more appropriate and human-like responses.

1 Introduction
Dialogue systems have attracted a lot of attention recently (Huang, Zhu, and Gao 2019). Sequence-to-sequence models (Sutskever, Vinyals, and Le 2014; Lei et al. 2018) are an effective framework that is commonly adopted in existing studies. However, a problem of sequence-to-sequence based methods is that they tend to generate generic and non-informative responses which provide deficient information (Gao et al. 2019).

Previous research has proposed various methods to alleviate the issue, such as adjusting objective functions (Li et al. 2016; Jiang et al. 2019), incorporating external knowledge (Ghazvininejad et al. 2018; Parthasarathi and Pineau 2018; Dinan et al. 2019), etc. Recently, Background Based Conversations (BBCs) have been proposed for generating more informative responses that are grounded in some background information (Zhou, Prabhumoye, and Black 2018; Moghe et al. 2018). As shown in Fig. 1, unlike previous conversational settings (Serban et al. 2016), in a BBC background material (e.g., a plot or review about a movie) is supplied to promote topic-specific conversations.

Figure 1: Background Based Conversation (BBC).

Existing methods for BBCs can be grouped into two categories, generation-based methods (e.g., GTTP (See, Liu, and Manning 2017)) and extraction-based methods (e.g., QANet (Yu et al. 2018)). Generation-based methods generate the response token by token, so they can generate natural and fluent responses, generally. However, generation-based methods suffer from two issues. First, they are relatively ineffective in leveraging background information. For example, for the case in Fig. 1, S2SA does not leverage background information at all. Second, they have difficulties locating the right semantic units in the background information. Here, a semantic unit is a span from the background information that expresses complete semantic meaning. For example, in Fig. 1, the background contains many semantic units, e.g., “mtv movie + tv awards 2004 best cameo” and “scary movie 4.” GTTP uses the wrong semantic unit “scary movie 4” to answer the question by “human 2.” Moreover, because generation-based methods generate the response one token at a time, they risk breaking a complete semantic unit, e.g., “scary movie 4” is split by a comma in the response of GTTP in Fig. 1. The reason is that generation-based methods lack a global perspective, i.e., each decoding step only focuses on a single (current) token and does not consider the tokens to be generated in the following steps. Extraction-based methods extract a span from the background as their response and are relatively good at locating

*Corresponding author
Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
the right semantic unit. But because of their extractive nature, they cannot generate natural conversational responses, see, e.g., the response of QANet in Fig. 1.

We propose a Reference-aware Network (RefNet) to address above issues. RefNet consists of four modules: a background encoder, a context encoder, a decoding switcher, and a hybrid decoder. The background encoder and context encoder encode the background and conversational context into representations, respectively. Then, at each decoding step, the decoding switcher decides between reference decoding and generation decoding. Based on the decision made by the decoding switcher, the hybrid decoder either selects a semantic unit from the background (reference decoding) or generates a token otherwise (generation decoding). In the latter case, the decoding switcher further determines whether the hybrid decoder should predict a token from the vocabulary or copy one from the background. Besides generating the response token by token, RefNet also provides an alternative way to learn to select a semantic unit from the background directly. Experiments on a BBC dataset show that RefNet significantly outperforms state-of-the-art methods in terms of both automatic and, especially, human evaluations.

Our contributions are as follows:

- We propose a novel architecture, RefNet, for BBCs by combining the advantages of extraction-based and generation-based methods. RefNet can generate more informative and appropriate responses while retaining fluency.
- We devise a decoding switcher and a hybrid decoder to adaptively coordinate between reference decoding and generation decoding.
- Experiments show that RefNet outperforms state-of-the-art models by a large margin in terms of both automatic and human evaluations.

## 2 Related work

We survey two types of related work on BBCs: generation-based and extraction-based methods.

### 2.1 Generation-based methods

Most effective generation-based models are based on sequence-to-sequence modeling (Sutskever, Vinyals, and Le 2014) and an attention mechanism (Bahdanau, Cho, and Bengio 2015). The proposed methods have achieved promising results on different conversational tasks (Serban et al. 2016). However, response informativeness is still an urgently need to be addressed challenge; these approaches prefer generating generic responses such as "I don’t know" and "thank you", which make conversations dull (Gao et al. 2019). Various methods have been proposed to improve response informativeness, such as adjusting objective functions (Li et al. 2016; Jiang et al. 2019), incorporating latent topic information (Xing et al. 2017), leveraging outside knowledge bases (Liu et al. 2018; Zhou et al. 2018) and knowledge representation (Ghazvininejad et al. 2018; Parthasarathi and Pineau 2018; Lian et al. 2019), etc.

Recently, Background Based Conversations (BBCs) have been proposed for generating more informative responses by exploring related background information (Zhou, Prabhumoye, and Black 2018; Dinan et al. 2019). Moghe et al. (2018) build a dataset for BBC and conduct experiments with state-of-the-art generation-based methods. They show that generation-based methods can generate fluent, natural responses, but have difficulty in locating the right background information. Therefore, most recent studies try to address this issue (Li et al. 2019; Qin et al. 2019). Zhang, Ren, and de Rijke (2019) introduce a pre-selection process that uses dynamic bi-directional attention to improve background information selection. Liu et al. (2019) propose an augmented knowledge graph based chatting model via transforming background information into knowledge graph. However, generation-based models still cannot solve inherent problems effectively, such as tending to break a complete semantic unit and generate shorter responses.

### 2.2 Extraction-based methods

Extraction-based methods have originally been proposed for Reading Comprehension (RC) tasks (Rajpurkar et al. 2016), where each question can be answered by a right span in a given passage. Wang and Jiang (2017) combine match-LSTM and a pointer network (Vinyals, Fortunato, and Jaitly 2015) to predict the boundary of the answer. Seo et al. (2016) propose BiDAF, which uses a variant co-attention architecture (Xiong, Zhong, and Socher 2017) to enhance the extraction result. Wang et al. (2017) propose R-Net, which introduces a self-matching mechanism. Yu et al. (2018) propose QANet, which devises an encoder consisting exclusively of convolution and self-attention. For BBCs, Moghe et al. (2018) show that extraction-based methods are better at locating the right background information than generation-based methods. However, current extraction-based methods are specifically designed for RC tasks. They are not suitable for BBCs for two reasons: First, BBCs usually do not have standard factoid questions like those in RC tasks. Second, BBCs require that the responses are fluent and conversational, which cannot be met by rigid extraction; see Fig. 1.

Unlike the work summarized above, we propose RefNet to combine the advantages of generation-based and extraction-based methods while avoiding their shortcomings. The main challenge that RefNet addresses is how to design an effective neural architecture that is able to refer to the right background information at the right time in the right place of a conversation while minimizing the influence on response fluency.

## 3 Reference-aware Network

Given a background in the form of free text \( K = (k_1, k_2, \ldots, k_n) \) with \( L_K \) tokens and a current conversational context \( C_\tau = (\ldots, X_{\tau-3}, X_{\tau-2}, X_{\tau-1}) \), the task of BBC is to generate a response \( X_\tau \) at \( \tau \).

Each \( X_\tau \) contains a sequence of \( L_X \) units, i.e., \( X_\tau = (x_{\tau}^{1}, x_{\tau}^{2}, \ldots, x_{\tau}^{L_X}) \), where \( x_{\tau}^{i} \) is the unit at timestamp \( t \) could be a token \( \{x_{\tau}^{i}\}_{i=1}^{n} \) or a semantic unit \( \{x_{\tau}^{i}\}_{i=1}^{n} \) containing \( n \) tokens.

RefNet consists of four modules: background encoder, context encoder, decoding switcher, and hybrid decoder; see
Fig. 2. Background and context encoders encode the given background \( \mathbf{H}^k \) and context \( \mathbf{C}_r \) into latent representations \( \mathbf{H}_c^k \) and \( \mathbf{H}_c^r \), respectively. \( \mathbf{H}_c^k \) and \( \mathbf{H}_c^r \) go through a matching layer to get a context-aware background representation \( \mathbf{H}_c^m \). At each decoding step, the decoding switcher predicts the probabilities of executing the reference decoding or generation decoding. The hybrid decoder takes \( \mathbf{H}_c^m \), \( \mathbf{H}_c^m \) and the embedding of the previous token as input and computes the probability of selecting a semantic unit from the background (reference decoding) or generating a token (generation decoding) based on the decision made by the decoding switcher. Next, we introduce the separate modules.

### 3.2 Hybrid decoder

During training, we know that the next \( x_t^r \) to be generated is a token \( \{x_{t,i}^r\}_{i=1}^n \) or a semantic unit \( \{x_{t,i}^c\}_{i=1}^n \). If \( x_t^r = \{x_{t,i}^c\}_{i=1}^n \), then \( x_t^r \) is generated in reference decoding mode with the probability modeled as follows:

\[
P(x_t^r | x_{<t}^r, C_r, K) = P(r)P(x_t^r | r),
\]

where \( P(r) \) is the reference decoding probability (see §3.3) and \( P(x_t^r | r) \) is the probability of generating \( x_t^r \) under the reference decoding. If \( x_t^r = \{x_{t,i}^c\}_{i=1}^n \), then \( x_t^c \) is generated in generation decoding mode with the probability modeled as:

\[
P(x_t^c | x_{<t}^c, C_r, K) =
\]

\[
P(g_p)P(x_t^c | g_p) + P(g_c)P(x_t^c | g_c),
\]

where \( P(g) = P(g_p) + P(g_c) \) is the decoding decoding probability (see §3.3) and \( P(g_p) \) is the predicting generation decoding probability (see §3.3). \( P(x_{t,i}^c | g_p) \) and \( P(x_{t,i}^c | g_c) \) are the probabilities of generating \( x_{t,i}^c \) under \( g_p \) and \( g_c \), respectively.

**Reference decoding.** Within reference decoding, the probability of generating the semantic unit \( \{x_{t,i}^c\}_{i=1}^n \) is evaluated as follows:

\[
P(x_t^c = \{x_{t,i}^c\}_{i=1}^n | r) = \alpha_{t, start}^r \alpha_{t, start+n-1}^r,
\]

where \( \alpha_{t, start}^r \) and \( \alpha_{t, start+n-1}^r \) are the probabilities of the start and end tokens of \( \{x_{t,i}^c\}_{i=1}^n \) (from the background), respectively, which are estimated by two-hop pointers with respect to the context-aware background hidden state sequence \( \mathbf{H}_c^m \). The \( \alpha_{t, start}^r \) is calculated by the first hop pointer, as shown in Eq. 7:

\[
\alpha_{t, start}^r = \frac{\exp(s_{t, start}^r)}{\sum_{j=1}^{L_{C_r}} \exp(s_{t,j}^r)},
\]

where \( s_{t,j}^r = v_{t,j}^r \text{tanh}(W_k^r h_j^c + U_k^r h_j^k + b_k^c) \), \( \alpha_{t,i}^c = \frac{\exp(s_{t,i}^c)}{\sum_{j=1}^{L_{C_r}} \exp(s_{t,j}^c)} \), and \( s_{t,i}^c = v_{t,i}^c \text{tanh}(W_{k,i}^c h_j^c + U_{k,i}^c h_j^k + b_{k,c}) \), with \( W_k^c, U_k^c, v_k^c \) and \( b_k^c \) being parameters.
where $W_{a_2}$, $W_r$, $U_r$, $v_r$, $b_{a_2}$ and $b_r$ are parameters. $h^x_t$ is the decoding hidden state vector, the updating scheme of which will be detailed in §3.4. $c^{se}_t$ and $c^{sm}_t$ are calculated in a similar way like Eq. 3 with $h^x_t$ attentively reading $H^c_t$ and $H^m$, respectively. The $\alpha^{r2}_{t,start+n-1}$ is calculated by the second hop pointer, as shown in Eq. 8:

$$c^r_t = \sum_{i=1}^{L_K} \alpha^r_{t,i} h^r_i + \alpha^o_t W_{a_2} [o^1_t; c^r_t] + b_{a_2},$$

$$s^r_{t,j} = v^r_t \tanh(W_{r} h^m_j + U_r o^2_t + b_r),$$

$$\alpha^{r2}_{t,start+n-1} = \frac{\exp(s^r_{t,j})}{\sum_{j=1}^{L_K} \exp(s^r_{t,j})},$$

where $W_{a_2}$ and $b_{a_2}$ are parameters. Reference decoding adopts soft pointers $\alpha^{r1}_{t,start}$ and $\alpha^{r2}_{t,start+n-1}$ to select semantic units, so it will not influence the automatic differentiation during training.

**Generation decoding.** Within predicting generation decoding, the probability of predicting the token $x^g_t$ from the vocabulary is estimated as follows:

$$P(x^g_t = x^g_{t,i} | g_p) = \text{softmax}(W_{g_p} o^1_t + b_{g_p}),$$

where $W_{g_p}$ and $b_{g_p}$ are parameters and the vector $o^1_t$ is the same one as in Eq. 7.

Within copying generation decoding, the probability of copying the token $x^c_t$ from the background is estimated as follows:

$$P(x^c_t = x_{c,i}^c | g_c) = \sum_{i:x^c_t = x^c_{i,j}} \alpha^{cm}_{t,i},$$

where $\alpha^{cm}_{t,i}$ is the attention probability distribution on $H^m$ produced by the same attention process with $c^{sm}_t$ in Eq. 7.

### 3.3 Decoding switcher

The decoding switching probabilities $P(r)$, $P(g_p)$ and $P(g_c)$ are estimated as follows:

$$[P(r), P(g_p), P(g_c)] = \text{softmax}(f_t),$$

where $f_t$ is a fusion vector, which is computed through a linear transformation in Eq. 12:

$$f_t = W_f [h^t; c^{se}_{t-1}; c^{sm}_{t-1}] + b_f,$$

where $W_f$ and $b_f$ are parameters. $h^t_t$ is decoding states (see §3.4).

During testing, at each decoding step, we first compute $P(r)$ and $P(g) = P(g_p) + P(g_c)$. If $P(r) \geq P(g)$, we use Eq. 4 to generate a semantic unit, otherwise we use Eq. 5 to generate a token.

### 3.4 State updating

The decoding state updating depends on whether the generated unit is a token or semantic unit. If $x^t_{t-1}$ is a token, then $h^t_t = \text{GRU}(h^t_{t-1}, [e(x^t_{t-1}); c^{se}_{t-1}; c^{sm}_{t-1}]).$

If $x^m_{t-1}$ is a span containing $n$ tokens, Eq. 13 will update $n$ times with one token as the input, and the last state will encode the full semantics of a span; see $h^t_t$ to $h^t_{t+n-1}$ in Fig. 2.

The decoding states are initialized using a linear layer with the last state of $H^m$ and $H^c$ as input:

$$h^0_0 = \text{ReLU}(W_{hs} [h^r_{L_K}; h^c_{L_c}] + b_{hs}),$$

where $W_{hs}$ and $b_{hs}$ are parameters. ReLU is the ReLU activation function.

### 3.5 Training

Our goal is to maximize the prediction probability of the target response given the context and background. We have three objectives, namely generation loss, reference loss and switcher loss.

The generation loss is defined as $L_g(\theta) = -\frac{1}{M} \sum_{\tau=1}^{M} \sum_{t=1}^{L_K} \log[P(x^g_t | x^c_{\tau,t}, C_r, K)],$ (15)

where $\theta$ are all the parameters of RefNet. $M$ is the number of all training samples given a background $K$. In $L_g(\theta)$, each $x^g_t$ is a token $\{x^g_{t,i}\}_{i=1}^{1}$. The reference loss is defined as $L_r(\theta) = -\frac{1}{M} \sum_{\tau=1}^{M} \sum_{t=1}^{L_K} I(x^c_t) \cdot \log[P(x^c_t | x^c_{\tau,t}, C_r, K)],$ (16)

where $I(x^c_t)$ is an indicator function that equals 1 if $x^c_t = \{x^c_{t,i}\}_{i=1}^{1}$ and 0 otherwise.

RefNet introduces a decoding switcher to decide between reference decoding and generation decoding. To better supervise this process we define switcher loss $L_s(\theta) = -\frac{1}{M} \sum_{\tau=1}^{M} \sum_{t=1}^{L_K} I(x^c_t) \log[P(r)] + (1 - I(x^c_t)) \log[P(g)],$ (17)

where $I(x^c_t)$ is also an indicator function, which is the same as in $L_s(\theta)$.

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

$$L(\theta) = L_g(\theta) + L_r(\theta) + L_s(\theta).$$ (18)

All parameters of RefNet as well as word embeddings are learned in an end-to-end back-propagation training paradigm.

### 4 Experimental Setup

#### 4.1 Implementation details

We set the word embedding size and GRU hidden state size to 128 and 256, respectively. The vocabulary size is limited to 25,000. For fair comparison, all models use the same embedding size, hidden state size and vocabulary size. Following Moghe et al. (2018), we limit the context length of all models to 65. We train all models for 30 epochs and test on a validation set after each epoch, and select the best model based on the validation results according to BLEU metric.
We use gradient clipping with a maximum gradient norm of 2. We use the Adam optimizer with a mini-batch size of 32. The learning rate is 0.001. The code is available online.\footnote{https://github.com/ChuanMeng/RefNet}

### 4.2 Dataset

Recently, some datasets for BBCs have been released (Zhou, Prabhumoye, and Black 2018; Dinan et al. 2019). We choose the Holl-E dataset released by Moghe et al. (2018) because it contains boundary annotations of the background information used for each response. We did not use other datasets because they do not have such annotations for training RefNet. Holl-E is built for movie chats in which each response is explicitly generated by copying and/or modifying sentences from the background. The background consists of plots, comments and reviews about movies collected from different websites. We use the mixed-short background which is truncated to 256 words, because it is more challenging according to Moghe et al. (2018). We follow the original data split for training, validation and test. There are also two versions of the test set: one with single golden reference (SR) and the other with multiple golden references (MR); see (Moghe et al. 2018).

### 4.3 Baselines

We compare with all methods we can get on this task.

- **Extraction-based methods**: (i) BiDAF extracts a span from background as response and uses a co-attention architecture to improve the span finding accuracy (Seo et al. 2016). (ii) R-Net proposes gated attention-based recurrent networks and a self-matching attention mechanism to encode background (Wang et al. 2017). (iii) QANet uses an encoder consisting exclusively of convolution and self-attention to capture local and global interactions in background (Yu et al. 2018).

- **Generation-based methods**: (i) S2S maps the context to the response with an encoder-decoder framework (Sutskever, Vinyals, and Le 2014). (ii) HRED encodes the context of the conversation with two hierarchical levels (Serban et al. 2016). S2S and HRED do not use any background information. (iii) S2SA adds an attention mechanism to the original S2S model to attend to the relevant background information (Bahdanau, Cho, and Bengio 2015). (iv) GTTP leverages background information with a copying mechanism to copy a token from the background at the appropriate decoding step (See, Liu, and Manning 2017). (v) CaKe is a improved version of GTTP, which introduces a pre-selection process that uses dynamic bi-directional attention to improve knowledge selection from background (Zhang, Ren, and de Rijke 2019). (vi) AKGCM first transforms background information into knowledge graph, and uses a policy network to select knowledge with an additional GTTP to generate responses (Liu et al. 2019).

\footnote{For fair comparison, different from Moghe et al. (2018), we do not use pre-trained GloVe (Pennington, Socher, and Manning 2014) such that all models randomly initialize the word embedding with the same vocabulary size.}

### 4.4 Evaluation metrics

Following the work of Moghe et al. (2018), we use BLEU-4, ROUGE-1, ROUGE-2 and ROUGE-L as automatic evaluation metrics. We also report the average length of responses outputted by each model. For extraction-based methods and RefNet, we further report F1 (Seo et al. 2016), which only evaluates the extracted spans not the whole responses. We also randomly sample 500 test samples to conduct human evaluations using Amazon Mechanical Turk. For each sample, we ask 3 workers to annotate whether the response is good in terms of four aspects: (1) **Naturalness** (N), i.e., whether the responses are 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.

### 5 Results

#### 5.1 Automatic evaluation

We list the results of all methods on mixed-short background setting in Table 1.

First, RefNet significantly outperforms all generation-based methods on all metrics, except in the BLEU score compared to AKGCM. Especially, RefNet outperforms the recent and strong baseline CaKe by around 2%-4% (significantly). The improvements show that RefNet is much better at leveraging and locating the right background information to improve responses than these generation-based methods. We believe RefNet benefits from reference decoding to tend to produce more complete semantic units, alleviating the inherent problems that pure generation-based method faced.

Second, RefNet outperforms extraction-based methods in most cases, including the strong baseline QANet. We think the reason is that extraction-based methods can only rigidly extract the relevant spans from the background, which does not consider the conversational characteristics of responses. Differently, RefNet also benefits from the generation decoding to generate natural conversational words in responses, which makes up the shortcoming of only extraction. RefNet is comparable in average length with extraction-based methods, which demonstrates that RefNet retains the advantages of extraction-based methods.

Third, the performance of these three extraction-based methods are comparable. However, their performances differ greatly between each other on the RC task dataset SQuAD (Rajpurkar et al. 2016), e.g., QANet outperforms BiDAF by around 7% on F1 score. Even with a stronger extraction-based model, we will arrive at a similar conclusion that they cannot generate natural and fluent responses due to the extraction nature. This confirms that extraction-based methods are not suitable for this task. Besides, we can further enhance the reference decoding of RefNet by incorporating various mechanisms used by extraction-based models. But that’s beyond the scope of this paper.
Table 1: Automatic evaluation results. **Bold face** indicates leading results. Significant improvements over the best baseline results are marked with * (t-test, \( p < 0.05 \)). SR and MR refer to test sets with single and multiple references. The results of AKGCM are taken from the paper because the authors have not released their code and processed knowledge graph. Note that AKGCM uses GloVe and BERT (Devlin et al. 2019) to improve performance but none of other models do.

<table>
<thead>
<tr>
<th>Methods</th>
<th>F1</th>
<th>BLEU</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-L</th>
<th>Average length</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SR</td>
<td>MR</td>
<td>SR</td>
<td>MR</td>
<td>SR</td>
<td>MR</td>
</tr>
<tr>
<td>no background</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S2S</td>
<td>-</td>
<td>-</td>
<td>5.26</td>
<td>7.11</td>
<td>27.15</td>
<td>30.91</td>
</tr>
<tr>
<td>HRED</td>
<td>-</td>
<td>-</td>
<td>5.23</td>
<td>5.38</td>
<td>24.55</td>
<td>25.38</td>
</tr>
<tr>
<td>mixed-short background (256 words)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiDAF</td>
<td>40.38</td>
<td>45.86</td>
<td>27.44</td>
<td>33.40</td>
<td>38.79</td>
<td>43.93</td>
</tr>
<tr>
<td>R-Net</td>
<td>40.92</td>
<td>46.84</td>
<td>27.54</td>
<td>33.18</td>
<td>39.78</td>
<td>44.30</td>
</tr>
<tr>
<td>QANet</td>
<td>41.65</td>
<td>47.32</td>
<td>28.21</td>
<td>33.91</td>
<td>40.66</td>
<td>44.82</td>
</tr>
<tr>
<td>GTTP</td>
<td>-</td>
<td>-</td>
<td>13.65</td>
<td>19.49</td>
<td>30.77</td>
<td>36.06</td>
</tr>
<tr>
<td>CaKe</td>
<td>-</td>
<td>-</td>
<td>26.03</td>
<td>29.18</td>
<td>40.21</td>
<td>44.12</td>
</tr>
<tr>
<td>AKGCM</td>
<td>-</td>
<td>-</td>
<td><strong>30.84</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RefNet</td>
<td><strong>41.86</strong></td>
<td><strong>48.46</strong></td>
<td><strong>30.33</strong></td>
<td><strong>33.97</strong></td>
<td><strong>42.11</strong></td>
<td><strong>47.35</strong></td>
</tr>
</tbody>
</table>

Table 2: Human evaluation results on mixed-short background (256 words). \( n \) 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)**.

<table>
<thead>
<tr>
<th>CaKe</th>
<th>QANet</th>
<th>RefNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \geq 1 )</td>
<td>( \geq 2 )</td>
<td>( \geq 1 )</td>
</tr>
<tr>
<td>(N)</td>
<td>449</td>
<td>264</td>
</tr>
<tr>
<td>(I)</td>
<td>359</td>
<td>115</td>
</tr>
<tr>
<td>(A)</td>
<td>390</td>
<td>153</td>
</tr>
<tr>
<td>(H)</td>
<td>438</td>
<td>231</td>
</tr>
</tbody>
</table>

5.2 Human evaluation

We also conduct a human evaluation for RefNet, CaKe (the best generation-based baseline), and QANet (the best extraction-based baseline). The results are shown in Table 2. Generally, RefNet achieves the best performance in terms of all metrics. In particular, we find that RefNet is even better than CaKe in terms of **Naturalness** and **Humanness**. We believe this is because RefNet has a good trade-off between **reference decoding** and **generation decoding**, where the generated conversational words and the selected semantic units are synthesized in a natural and appropriate way. RefNet is also much better than CaKe in terms of **Informativeness** and **Appropriateness**, which shows that RefNet is better at locating the appropriate semantic units. The reason is that with the ability to generate a full semantic unit at once, RefNet has a global perspective to locate the appropriate semantic units, reducing the risk of breaking a complete semantic unit. QANet achieves good evaluation scores on **Informativeness** and **Appropriateness** than CaKe, but gets the worst scores on **Naturalness** and **Humanness**. Although QANet is relatively good at locating the relevant semantic unit, its responses lack contextual explanations, which makes workers hard to understand. This further shows that only extracting a span from the background is far from enough for BBCs, even replacing QANet with a more stronger extraction-based one.

6 Analysis

6.1 Reference vs. generation decoding

We analyze the effectiveness of reference and generation decoding, we compare the results of RefNet with only reference decoding (force reference) and with only generation decoding (force generation) in Table 3. Note that force generation is better than GTTP because there are two differences. First, we use a matching layer to get the context-aware background representation in Eq. 2, while GTTP only uses basic background representations without such a matching operation. Second, we use the hidden states of the background and context to jointly initialize the decoding states in Equation. Second, we use the hidden states of the background representations without such a matching operation. RefNet is better than GTTP because there are two differences.

We use the code released by Moghe et al. (2018) https://github.com/nikitacs16/Holl-E
Table 4: Case study. Bold face indicates the true span in the current turn.

<table>
<thead>
<tr>
<th>Example 1</th>
<th>Example 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Background:</strong> ... but if you like ben stiller, go see ” meet the fockers “, dustin’s antics will favorite character was jack ( the older one ) , because he was so serious but always plotting and putting up a front. i think it was $279,167,575 awards ascap film and television music awards 2005 top box office films mtv ...</td>
<td><strong>Background:</strong> ... being captured by boris and onatopp. bond arrives in st. petersburg and meets his cia contact, jack wade ( joe don baker ). wade agrees to take bond to the hide-out of a russian gangster, valentin lukovsky ( robbie coltrane ) , whom bond had shot in the leg and given a permanent limp years before ...</td>
</tr>
<tr>
<td>Human1: that name is so ridiculous but funny. Human2: first off , the writers did not miss a single opportunity to play off of the name ” focker “. Human1: yeah, i heard it was a pretty successful movie overall.</td>
<td>Human1: that was a good seen. Human2: what did you like about the movie? Human1: i liked his friend, jack wade.</td>
</tr>
<tr>
<td>CaKe</td>
<td>QANet $279,167,575</td>
</tr>
<tr>
<td>RefNet</td>
<td>i made $279,167,575 at the box office.</td>
</tr>
</tbody>
</table>

Table 5: Analysis of switcher loss on mixed-short background version. Bold face indicates leading results. Significant improvements over the best competitor are marked with * ($t$-test, $p < 0.05$).

<table>
<thead>
<tr>
<th>Without SL</th>
<th>With SL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SR</td>
</tr>
<tr>
<td>F1</td>
<td>37.13</td>
</tr>
<tr>
<td>BLEU</td>
<td>28.96</td>
</tr>
<tr>
<td>ROUGE-1</td>
<td>41.27</td>
</tr>
<tr>
<td>ROUGE-2</td>
<td>30.65</td>
</tr>
<tr>
<td>ROUGE-L</td>
<td>36.02</td>
</tr>
</tbody>
</table>

6.2 Switcher loss

To verify the effectiveness of the switcher loss $L_s(\theta)$ in Eq. 14, while GTTP only uses the single representation of background to initialize it. We can see that force reference and force generation are comparable if working alone. The contributions of reference and generation decoding are complementary as the combination brings further improvements on all metrics, demonstrating the need for both.

6.3 Case study

We select some examples from the test set to illustrate the performance of RefNet, CaKe and QANet, as shown in Table 4. One can see that RefNet can select the right semantic unit from the background or generate fluent tokens at appropriate time and position, resulting in more informative and appropriate responses. For instance, in Example 1, RefNet identifies the right semantic unit “$279,167,575” within the background, which is combined with “it made” ahead and followed by “at the box office” to form a more natural and conversational response. The second example indicates that RefNet can locate longer semantic units accurately. In contrast, the responses by QANet lack naturality. The responses by CaKe are relatively inconsistent and irrelevant. In the first example, CaKe breaks the complete semantic unit “if you like ben stiller” and throws out the part “if you like”.

There are also some cases where RefNet does not perform well. For example, we find that RefNet occasionally selects short or meaningless semantic units, such as “i” and “it.” This indicates that we could further improve reference decoding by taking more factors (e.g., the length of semantic units) into consideration.

7 Conclusion and Future Work

In this paper, we propose RefNet for the Background Based Conversation (BBCs) task. RefNet incorporates a novel reference decoding module to generate more informative responses while retaining the naturality and fluency of responses. Experiments show that RefNet outperforms state-of-the-art methods by a large margin in terms of both automatic and human evaluations.

A limitation of RefNet is that it needs boundary annotations of semantic units to enable supervised training. In future work, we hope to design a weakly supervised or unsupervised training scheme for RefNet in order to apply it to other datasets and tasks. In addition, we will consider more factors (e.g., the length or frequency of semantic unit) to further improve the reference decoding module of RefNet.
Acknowledgments
We thank the anonymous reviewers for their helpful comments. This work is supported by the Natural Science Foundation of China (61972234, 61902219, 61672324, 61672322), the Natural Science Foundation of Shandong province (2016ZRE27468), the Tencent AI Lab Rhino-Bird Focused Research Program (JR201932), the Fundamental Research Funds of Shandong University, Ahold Delhaize, the Association of Universities in the Netherlands (VSNU), and the Innovation Center for Artificial Intelligence (ICAI).

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