Task-aware response generation in task-oriented dialogue with unstructured knowledge access

Abstract

To alleviate the problem of structured databases’ limited coverage, recent task-oriented dialogue systems incorporate external unstructured knowledge to guide the generation of system responses. However, these usually use word or sentence level similarities to detect the relevant knowledge context, which only partially captures the top level relevance. In this paper, we examine how to better integrate topical information in knowledge grounded task-oriented dialogue and propose “Topic-Aware Response Generation” (TARG), an end-to-end response generation model. TARG incorporates multiple topic-aware attention mechanisms to derive the importance weighting scheme over dialogue utterances and external knowledge sources towards a better understanding of the dialogue history. Experimental results indicate that TARG achieves state-of-the-art performance in knowledge selection and response generation, outperforming previous state-of-the-art by 3.2, 3.6, and 4.2 points in EM, F1 and BLEU-4 respectively on Doc2Dial, and performing comparably with previous work on DSTC9; both being knowledge-grounded task-oriented dialogue datasets.

1 Introduction

Task-oriented (or goal-oriented) dialogue systems aim to accomplish a particular task (e.g. book a table, provide information) through natural language conversation with a user. The system’s available actions are often described by a pre-defined domain-specific schema while relevant knowledge is retrieved from structured databases or APIs (Rastogi et al., 2020). As such, task-oriented dialogue systems are often limited on which actions can be taken and what information can be retrieved (Kim et al., 2020). To relax these restrictions, some dialogue systems (also referred to as goal-oriented chatbots) adopt open-domain language that is by definition unconstrained by pre-defined actions (Feng et al., 2020), and dynamically extract any required knowledge from in-domain unstructured collections in the form of entity descriptions, FAQs, and documents. Access to external knowledge sources has also been shown to help dialogue systems generate more specific and informative responses, which helps with the “common response” problem (Zhang et al., 2018; Ren et al., 2020).

Figure 1 shows an example of a task-oriented dialogue that exploits external unstructured knowledge sources. Given a history of previous dialogue turns, with each turn consisting of one user and system utterance, and access to in-domain unstructured knowledge sources (either a document collection or a set of candidate facts), the dialogue system needs to generate an appropriate system response for the current turn. Recent research (Zhang et al., 2018; Ren et al., 2020) tackles the task by decomposing it into two sub-tasks: to initially determine the relevant knowledge (if any) that needs to be extracted/selected from external resources, and to subsequently generate the response based on the selected knowledge and the dialogue history.

When retrieving knowledge from unstructured sources, different sources may need to be accessed
in different dialogue turns; this is to be expected in most conversation scenarios. In the example of Figure 1, the first turn is grounded on the first knowledge candidate, and subsequent turns are grounded on later candidates. If we consider that each knowledge source belongs to a different topic or domain (e.g., "how you apply", "publications", "application is denied" in our example), we can observe that as the knowledge selection shifts across sources during the course of the dialogue, a corresponding shift occurs between topics. Previous work has not actively exploited this, but we posit that attending the topic shifts in the dialogue history can provide signals that help distinguish relevant from irrelevant sources for knowledge selection, and that such topical information can help the model derive an importance weighting scheme over the dialogue history for better response generation.

In this paper, we model topic shifts in selected knowledge sources to improve topic-aware knowledge selection and response generation in task-oriented dialogue, and propose “Topic-Aware Response Generation” (TARG), an end-to-end model for knowledge selection and response generation. Our approach incorporates multiple topic-aware attention mechanisms to derive the importance weighting scheme over previous utterances and knowledge sources, aiming for a better understanding of the dialogue history. In addition, TARG is built on top of recent breakthroughs in language representation learning by finetuning on the pretrained language model BART (Lewis et al., 2020).

We conduct extensive experiments with two task-oriented dialogue datasets, namely Doc2Dial (Feng et al., 2020) and DSTC9 (Gunasekara et al., 2020). Our results indicate that TARG is able to accurately select the appropriate knowledge source, and as a result generate more relevant and fluent responses, outperforming previous state-of-the-art by 3.2, 3.6, and 4.2 points in EM, F1 and BLEU-4 respectively on Doc2Dial, and performing comparably with previous work on DSTC9. Furthermore, we present an ablation study and a case study accompanied by analysis of the learned attention mechanisms.

2 Related Work

As we briefly mentioned in the introduction, the majority of previous work decomposed knowledge-grounded dialogue generation into two sub-tasks: knowledge selection and response generation.

To determine the relevant candidate for knowledge selection, the use of keyword matching (Ghazvininejad et al., 2018), information retrieval (Young et al., 2018) and entity diffusion (Liu et al., 2018) methods have been proposed. More specifically, keyword matching methods (Bordes et al., 2017) focus on calculating a weight for each keyword in the knowledge candidate and then determine their relevance based on the weighted sum of the keywords’ representations. On the other hand, some information retrieval techniques compute traditional tf-idf scores to detect the knowledge candidate in the most relevant document to the user’s query (Song et al., 2018; Dinan et al., 2018), while others leverage the power of neural networks to learn a candidate ranking function directly through an end-to-end learning process (Yan and Zhao, 2018; Zhao et al., 2019; Gu et al., 2020). Another approach uses entity diffusion networks (Wang et al., 2020) that perform fact matching and knowledge diffusion to ground both knowledge candidates and dialogues.

For response generation, the related work has adapted both response retrieval and language generation approaches. Specifically for response retrieval, deep interaction networks (Sun et al., 2020) have been employed to learn better-suited representations to ground candidate responses against external knowledge, while language generation approaches have been adapted to attend to ground knowledge during inference (Peng et al., 2020), with some further employing copy mechanisms over both dialogue context and external knowledge (Yavuz et al., 2019), or leveraging a reading comprehension model to similarly extract relevant spans (Qin et al., 2019; Wu et al., 2021).

Recently, pre-trained language models such as BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019), which have demonstrated significant improvements on numerous natural language processing tasks, have also been applied to improve model the semantic representation in knowledge selection and response generation (Zhao et al., 2020; Li et al., 2020; Feng et al., 2020). Alternatively, other approaches combine the generative capability of autoregressive decoders such as GPT-2 (Budzianowski and Vulić, 2019) or T5 (Raffel et al., 2020), to better generate the system response.

Broader dialogue research has explored the topic-aware signal present in the dialogue history, but such work did not consider external knowledge nor its topics. Briefly, Xing et al. (2017) proposed a
topic-aware seq-to-seq approach for open-domain dialogue that attends over LDA topics inferred from the dialogue history, while Zhang et al. (2020) calculates the relevance between topic distributions of the dialogue history and the immediate context and attends over them to generate the next system response. In retrieval-based dialogue systems, Xu et al. (2021b) performs topic-aware segmentation of the context to better inform dialogue modeling.

We briefly discuss more recent work in our experiments section, as we compare it against our approach. To the best of our knowledge no other work has explicitly modelled the topic shifts in both dialogue history and external knowledge to inform knowledge selection and response generation in knowledge-ground task-oriented dialogue systems.

3 Our Approach

As we mentioned in the introduction, our proposed approach (TARG) exploits topic-aware mechanisms to derive an importance weighting scheme over different utterances in the dialogue history, with the goal to better inform knowledge selection and response generation. For a brief overview of TARG, please consult Figure 2. The input in our task consists of the dialogue history of previous user and system utterances, and a set of external knowledge candidates (hereafter referred to as factoids for brevity). The goal is to generate the next system utterance in the dialogue, which may or may not be grounded in one of the factoids; some of the dialogue history utterances may also be grounded on factoids but not necessarily all of them are.

Briefly, to generate the next turn’s system utterance, TARG initially generates BART-based representations for every previous user and system utterance in the dialogue history, for every available factoid, and for both utterances’ and factoids’ corresponding topics. For each utterance / factoid pair, TARG extracts matching features by calculating feature interaction over their encoded representations. TARG subsequently weights the matching features by topic-aware attention mechanisms, and aggregates them in a tensor. Finally, a knowledge selection layer outputs a relevance score over factoids, and the decoder generates the system utterance based on the most relevant factoid’s encoding.

3.1 Utterance and Factoid Encoder

We use a BART encoder to generate representations for every utterance in the dialogue history (up to a maximum history length) and factoid in external knowledge. We similarly, but separately, generate representations for their corresponding topics. Our work assumes that the corresponding topic of factoids can be derived in some way from the available data, e.g. the topic can be interpreted as the title of the factoid’s originating document or its annotated domain. While we do not explore the possibility in this paper, the topic could also potentially be inferred using topic modelling techniques. The topic of each utterance is considered the same as that of their corresponding factoids (if any). Since not all dialogue turns are necessarily grounded in external knowledge, in absence of a corresponding factoid, the topic is set to a generic “non-relevant” pseudo-topic. This process results in the semantics and topic of every utterance or factoid being represented explicitly by separate embeddings.

Specifically, in order to generate the semantic embeddings $s_u$ and $s_k$ of every utterance and factoid respectively, the token sequence $X = ([CLS], x_1, ..., x_N, [SEP], [MODE], [CLS])$ is passed through a BART encoder, where the subword tokens of the text are denoted as $x_1, ..., x_N$. [CLS] and [SEP] are start-of-text/end-of-text.
and separator pseudo-tokens respectively, while [MODE] is one of [SYS]/[USER]/[KLG] to indicate whether the text belongs to a system utterance, user utterance, or factoid respectively. The state of the final [CLS] is used as the utterance’s / factoid’s semantic embedding. Similarly, to generate the topic embeddings \( t_u \) and \( t_k \) of every utterances and factoid, the BART encoder sequence input is
\[
T = ([CLS], x_1, ..., x_N, [SEP], [MODE], [POSIT], [CLS]), \nonumber
\]
where [POSIT] is the position of the corresponding dialogue history utterance (zero if the text belongs to a factoid). The state of the final [CLS] is used as the topic embedding.

### 3.2 Topic-aware Attention

In the next step, TARG calculates feature interactions over the semantic embeddings to extract matching features, which are subsequently weighted by a number of topic-aware attention mechanisms. These attention mechanisms operate over the topic embeddings of utterances and factoids to calculate topic-aware utterance / factoid pair matching representations. The motivation is to incorporate a more flexible way to weight and aggregate matching features of different dialogue history utterances with topic-aware attention, so that the model learns to better attend over them.

Specifically, we design three different types of topic-aware attention that are calculated between each topic embedding \( t_k \), corresponding to the \( i \)-th factoid, and the topic embeddings of all utterances in dialogue history, as follows:

**Dot Product.** We concatenate the utterance topic embeddings \( t_u \) with the factoid topic embedding, and compute the dot product between parameter \( w_d \) and the resulting vector:
\[
A^i_d = \text{softmax}(w_d^T [t_u, t_k]) \quad \forall t_u \in T_u \quad (1)
\]

**Bilinear.** We compute the bilinear interaction between \( t_u \) and \( t_k \) and then normalize the result:
\[
A^i_b = \text{softmax}(w_b^T t_u t_k^T) \quad \forall t_u \in T_u \quad (2)
\]

where \( W_b \) is a bilinear interaction matrix.

**Outer Product.** We compute the outer product between \( t_u \) and \( t_k \), then project this feature vector through a fully connected layer and a softmax:
\[
A^i_o = \text{softmax}(w_o^T (t_u \times t_k)) \quad \forall t_u \in T_u \quad (3)
\]

where \( w_o \) is a parameter and \( \times \) is the outer product.

In parallel, we calculate the feature interaction matrix \( F_i \) between the semantic embeddings of all utterances \( s_u^{i} \) and the factoid \( s_k^{i} \). Every row \( F_{i,j} \) of \( F_i \) is calculated as follows:
\[
F_{i,j} = v_f^T \tanh(s_u^{i} W_f s_k^{jT} + b_f) \quad (4)
\]

with \( W_f, b_f, v_f \) being model parameters.

To obtain a unified utterance / factoid pair representation \( k_i \) for each factoid \( i \), we concatenate the weighted sums of all utterances / factoid interaction embeddings with the different attention mechanisms. The final topic-aware utterance / factoid pair representation across all factoids is \( K \), where the \( i \)-th column vector is \( k_i \):
\[
k_i = [A^T_d F_i, A^T_b F_i, A^T_o F_i] \quad (5)
\]

### 3.3 Relevant Knowledge Selection

For the purpose of knowledge selection, TARG treats all external knowledge as a single document, by simply concatenating all available factoids. To account for the possibility that the system response shouldn’t be grounded on any external knowledge, a “non-relevant” pseudo-factoid is included.

The relevant knowledge selector takes the topic-aware representations of these sequential factoids as input and predicts a span over the overall document that the system response should be grounded on. Through this process, several knowledge candidates may appear in the selected span.

The grounded span is derived by predicting the start and the end indices of the span in the document. We obtain the probability distribution of the start and the end indices of the span in the document by the following equations:

\[
p_s = \text{softmax}(W_s^T K + b_s), \quad (6)
\]

\[
p_e = \text{softmax}(W_e^T K + b_e), \quad (7)
\]

where \( W_s, W_e, b_s, b_e \) are trainable weight vectors.

### 3.4 System Response Generation

The system response generator decodes the response by attending on the selected knowledge span. Since the span may contain several factoids, we first use a Convolution Neural Network (CNN) to fuse the information. We apply this CNN even when only a single factoid is present in the span for consistency. The CNN receives the topic-aware
We evaluate our proposed approach on two where
leased dataset with a withheld test set used for the
Table 1: Number of dialogues, documents and average
number of content elements per document (tk: tokens,
sp: spans, p: paragraphs, sec: titled sections) per domain
in Doc2dial.

<table>
<thead>
<tr>
<th>Domain</th>
<th>#Dials</th>
<th>#Docs</th>
<th>avg # per doc</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>tk</td>
</tr>
<tr>
<td>ssa</td>
<td>1192</td>
<td>109</td>
<td>795</td>
</tr>
<tr>
<td>va</td>
<td>1330</td>
<td>138</td>
<td>818</td>
</tr>
<tr>
<td>dmv</td>
<td>1305</td>
<td>149</td>
<td>944</td>
</tr>
<tr>
<td>studentaid</td>
<td>966</td>
<td>91</td>
<td>1007</td>
</tr>
<tr>
<td>all</td>
<td>4793</td>
<td>487</td>
<td>888</td>
</tr>
</tbody>
</table>

Table 2: Number of dialogues, snippets and average
number of content elements per snippet (tk: tokens,
sent: sentences) per domain in the DSTC9 dataset.

<table>
<thead>
<tr>
<th>Domain</th>
<th>#Dials</th>
<th>#Snippets</th>
<th>#per-snippet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>tk</td>
</tr>
<tr>
<td>Hotel</td>
<td>-</td>
<td>1219</td>
<td>9</td>
</tr>
<tr>
<td>Restaurant</td>
<td>-</td>
<td>1650</td>
<td>7</td>
</tr>
<tr>
<td>Train</td>
<td>-</td>
<td>26</td>
<td>15</td>
</tr>
<tr>
<td>Taxi</td>
<td>-</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>all</td>
<td>10,438</td>
<td>2900</td>
<td>8</td>
</tr>
</tbody>
</table>

The DSTC9 dataset also includes conversation
dialogues, but the external knowledge is in the form
of FAQ documents, in essence containing question
answering pairs on a specific domain; we consider
each pair as a distinct factoid and their domain as
the topic. In practice, these FAQs are to be used to
answer follow-up user questions that are out of the
coverage of a dialogue system’s database. Similarly
to Doc2Dial, we observe that the focused “topic”
in the DSTC9 dataset is also varied throughout the
communications. Table 1 and Table 2 presents the
statistics of the Doc2dial and DSTC9 datasets.

3.5 Optimization
During training, we optimize both the knowledge
selector and response generator via their cross-
entropy losses $L_s, L_g$ respectively. We compute
the joint loss $L$ as follows:

$$L = \lambda \cdot L_s + (1 - \lambda) \cdot L_g,$$

where $\lambda \in [0, 1]$ is a balance coefficient.

4 Experiments
4.1 Datasets
We evaluate our proposed approach on two
benchmark data sets on task-oriented dialogue:
Doc2dial (Feng et al., 2020) and DSTC9 (Gus-
asekara et al., 2020). Doc2dial is a recently re-
leased dataset with a withheld test set used for the
corresponding leaderboard, which includes conver-
sation dialogues between an assisting system and
an end user, with an accompanying set of docu-
ments wherein distinct factoids are clearly anno-
tated; further annotations indicate which dialogue
utterances are grounded on which factoids of the
associated documents. The Doc2dial dataset includes
many cases of conversations that are grounded
on factoids from different documents. If we consider
the title of each document as a distinct topic, then
each of these conversations can be interpreted to
involve many interconnected topics under a general
inquiry, making it an ideal dataset for our approach.

The DSTC9 dataset also includes conversation
dialogues, but the external knowledge is in the form
of FAQ documents, in essence containing question
answering pairs on a specific domain; we consider
each pair as a distinct factoid and their domain as
the topic. In practice, these FAQs are to be used to
answer follow-up user questions that are outside the
coverage of a dialogue system’s database. Similarly
to Doc2Dial, we observe that the focused “topic”
in the DSTC9 dataset is also varied throughout the
communications. Table 1 and Table 2 present the
statistics of the Doc2dial and DSTC9 datasets.

4.2 Baselines
In the following experiments, we compare our ap-
proach against previously published state-of-the-art
approaches on the Doc2dial and DSTC9 datasets.
We have not re-implemented these approaches,
but report their already published results for the
datasets for which they are available.2

Doc2Dial-baseline (Feng et al., 2020): This is the
baseline provided by the Doc2Dial challenge. It
consists of an extractive question answering model

2While there are better performing systems in the DSTC9
and Doc2Dial leaderboards, these are either not published, not
based on a single method, or exploit additional external data,
and thus are not directly comparable to this work.
Table 3: Performance of TARG and related work on the DSTC9 dataset; baseline refers to the DSTC9 provided baseline. Numbers in bold denote best results in that metric.

<table>
<thead>
<tr>
<th>Model</th>
<th>Knowledge Selection</th>
<th>Response Generation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MRR@5</td>
<td>Recall@5</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.772</td>
<td>0.877</td>
</tr>
<tr>
<td>KDEAK</td>
<td>0.853</td>
<td>0.896</td>
</tr>
<tr>
<td>RADGE</td>
<td><strong>0.937</strong></td>
<td>0.966</td>
</tr>
<tr>
<td>EGR</td>
<td>0.894</td>
<td>0.934</td>
</tr>
<tr>
<td>TARG</td>
<td>0.935</td>
<td>0.972</td>
</tr>
</tbody>
</table>

Table 4: Performance of TARG and related work on the Doc2Dial dataset; baseline refers to the Doc2Dial provided baseline. Numbers in bold denote best results in that metric.

<table>
<thead>
<tr>
<th>Model</th>
<th>Knowledge Selection</th>
<th>Response Generation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EM</td>
<td>F1</td>
</tr>
<tr>
<td>Baseline</td>
<td>37.2</td>
<td>52.9</td>
</tr>
<tr>
<td>JARS</td>
<td>42.1</td>
<td>57.8</td>
</tr>
<tr>
<td>CAiRE</td>
<td>45.7</td>
<td>60.1</td>
</tr>
<tr>
<td>RWTH</td>
<td>46.6</td>
<td>62.8</td>
</tr>
<tr>
<td>TARG</td>
<td><strong>49.8</strong></td>
<td><strong>66.4</strong></td>
</tr>
</tbody>
</table>

4.3 Evaluation Measures

We make use of the following automatic evaluation metrics in our experiments. For each dataset, we calculate the metrics used by the respective challenges for consistency.

- **Exact Match (EM)**: This measures what part of the predicted knowledge span matches the ground truth factoid exactly.
- **Token-Level F1**: We cast the predicted spans and ground truth factoids as bags of tokens, and compute F1 between them.
- **MRR@5**: A metric based on the rank of the first ground truth factoid in a system’s top-5 ranking.
- **Recall@5**: This metric counts how many ground truth factoids occur in a system’s top-5 ranking.
- **BLEU-X** (Papineni et al., 2002): BLEU-X estimates a generated response’s via measuring its n-gram precision against the ground truth. X denotes the maximum size of the considered n-grams (i.e. unigrams, bigrams, trigrams, 4-grams).
- **ROUGE-X** (Lin, 2004): ROUGE-X measures n-gram recall between generated and ground truth response. ROUGE-L measures the longest common word subsequence.

4.4 Implementation Details

We use a pre-trained BART-base model to encode utterances and factoids. The max sentence length is set to 50 and the max number of dialogue turns is set to 15. The hidden size of attentions are all

...
set to 768. The size of the convolution and pooling kernels are set to (3, 3, 3). The joint loss $\lambda$ is 0.5. The dropout probability is 0.1. The batch size is set to 8. We optimize with Adam and an initial learning rate of $3\times 10^{-5}$.

### 4.5 Experimental Results

Table 3 and Tables 4 show our results on DSTC9 and Doc2Dial respectively. Observe that TARG performs significantly better than related work in both knowledge selection and response generation on the Doc2Dial dataset, outperforming the second best system by 3.2, 3.6, and 4.2 points in EM, F1 and BLEU-4 respectively.

On the DSTC9 dataset, TARG outperforms the related work in most metrics, though by narrow margins. Due to the smaller differences, we consider TARG to be performing on par with state-of-the-art on DSTC9. The performance gains of TARG can be explained by the topic-aware mechanism as it provides a more flexible way to weight and aggregate different dialogue history turns. This indicates that better understanding of the dialogue history is crucial for predicting the relevant factoids and generating a reasonable response.

### 5 Discussion

#### 5.1 Ablation Study

Here we conduct an ablation study of TARG, to explore the effects of the BART model, topic-aware attention, as well as the different topic attention mechanisms. The results indicate that all these mechanisms are necessary to the performance of knowledge selection and response generation.

#### Effect of BART: To investigate the effectiveness of using BART in the utterance / factoid encoder and system response generator, we replace BART with a bi-directional LSTM and rerun the model for Doc2Dial and DSTC9. As shown in Figures 3 and 4, the performance of the BiLSTM-based model TARG-w/oBART decreases significantly in knowledge selection, and especially in response generation as is indicated by the drop in BLEU. As expected, this indicates that the BART model can create and utilize more accurate representations for dialogue history and unstructured knowledge.

#### Effect of topic-aware attention: Next we remove the topic-aware attention mechanisms (TARG-w/oAtt). Figures 3 and 4 again show that the respective performances deteriorate considerably. This shows that topic-aware attention helps derive an important weighting scheme over the utterances leading to better understanding of dialogue history.

#### Effect of topic attention mechanisms: Here we compare TARG against TARG-dot, TARG-bilinear, and TARG-outer which use exclusively doc product attention, bilinear attention, and outer product attention respectively. Table 5 shows that doc product attention underperforms compared to bilinear and outer product attention while bilinear attention’s performance is comparable with outer product attention. In addition, any isolated attention mechanism performs considerably worse than their fusion, supporting its utilization. We conjecture that this is due to how different attention mechanisms focus on different topic features.

#### 5.2 Case Study

Consult Figure 5 for a case study from the Doc2Dial dataset. On the top of the Figure are the previous turns of dialogue history, while on the right is a subset of the available factoids. We can observe how the topic changes throughout the turns of dialogue history (by consulting the corresponding factoid topic), from “Exploring Your
In this paper, we proposed TARG: “Topic-Aware Response Generation”, a topic-aware model which incorporates multiple topic-aware attention mechanisms to derive the importance weighting scheme over both dialogue utterances and unstructured external knowledge, and through that facilitate better dialogue history understanding. Our proposed method achieves state-of-the-art results in both knowledge selection and response generation, outperforming previous state-of-the-art by 3.2, 3.6, and 4.2 points in EM, F1 and BLEU-4 respectively on Doc2Dial, and performing comparably with previous work on DSTC9. To provide further insights, we also presented an ablation study of our model that supported the importance of our method’s various components, and discussed a case study accompanied by an analysis of the attention mechanisms.

6 Conclusion
References


Sopan Khosla, Justin Love lace, Ritam Dutt, and Adithya Pratapa. 2021. Team jars: Dialdoc subtask 1-improved knowledge identification with supervised out-of-domain pretraining. In ACL.


Linxiao Li, Can Xu, Wei Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. Zero-resource knowledge-grounded dialogue generation. In NIPS.


