A Knowledge-Grounded Multimodal Search-Based Conversational Agent

Shubham Agarwal *, Ondřej Dušek, Ioannis Konstas and Verena Rieser
The Interaction Lab, Department of Computer Science
Heriot-Watt University, Edinburgh, UK
* Adeptmind Scholar, Adeptmind Inc., Toronto, Canada
{sa201, o.dusek, i.konstas, v.t.rieser}@hw.ac.uk

Abstract

Multimodal search-based dialogue is a challenging new task: It extends visually grounded question answering systems into multi-turn conversations with access to an external database. We address this new challenge by learning a neural response generation system from the recently released Multimodal Dialogue (MMD) dataset (Saha et al., 2017). We introduce a knowledge-grounded multimodal conversational model where an encoded knowledge base (KB) representation is appended to the decoder input. Our model substantially outperforms strong baselines in terms of text-based similarity measures (over 9 BLEU points, 3 of which are solely due to the use of additional information from the KB).

1 Introduction

Conversational agents have become ubiquitous, with variants ranging from open-domain conversational chit-chat bots (Ram et al., 2018; Papaioannou et al., 2017; Fang et al., 2017) to domain-specific task-based dialogue systems (Singh et al., 2000; Rieser and Lemon, 2010, 2011; Young et al., 2013; Wen et al., 2017).

Our work builds upon the recently released Multimodal Dialogue (MMD) dataset (Saha et al., 2017), which contains dialogue sessions in the e-commerce (fashion) domain. Figure 1 illustrates an example chat session with multimodal interaction between the user and the system. We focus on the task of generating textual responses conditioned on the previous conversational history. Traditional goal-oriented dialogue systems relied on slot-filling approach to this task, i.e. explicit modelling of all attributes in the domain (Lemon et al., 2006; Wang and Lemon, 2013; Young et al., 2013). On the other hand, previous work on MMD data used direct learning from raw texts with implicit semantic representation only. This paper attempts to combine both approaches by learning to generate replies from raw user input, while also incorporating Knowledge Base (KB) inputs (i.e. explicit semantics) into the generation process. We discuss how our model is able to handle various user intents (request types) and the impact of incorporating the additional explicit semantic information from the KB into particular targeted intents. We use user intent annotation and KB queries provided with the dataset for the purpose of this work.

Our main contribution is the resulting fully data-driven model for the task of conversational multimodal dialogue generation, grounded in conversational text history, vision and KB inputs. We also illustrate a method to improve context modelling over multiple images and show great improvements over the baseline. Finally, we present a detailed analysis of the outputs generated by our system corresponding to different user intents.

2 Related Work

With recent progress in deep learning, there is continued interest in the tasks involving both vision and language, such as image captioning (Xu et al., 2015; Vinyals et al., 2015; Karpathy and Fei-Fei, 2015), visual storytelling (Huang et al., 2016), video description (Venugopalan et al., 2015b,a) or dialogue grounded in visual context (Antol et al., 2015; Das et al., 2017; Tapaswi et al., 2016).

Bordes et al. (2016) and Ghazvininejad et al. (2017) presented knowledge-grounded neural models; however, these are uni-modal in nature, involve only textual interaction and do not take into account the conversational history in a dia-

Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd Int'l Workshop on Search-Oriented Conversational AI, pages 59–66
Brussels, Belgium, October 31, 2018.
ISBN 978-1-948087-75-9

U1 **SHOPPER:** Hi there

A1 **AGENT:** Hi, anything i can help you with today?

U2 **SHOPPER:** I am here to shop for tapered type casual trousers that my brother would like .

U3 **SHOPPER:** he is 26 years of age.

A2 **AGENT:** Oh. Good

A3 **AGENT:** let me just make a quick search through my catalogue

A4 **AGENT:**

U4 **SHOPPER:** Are the products in the 5th, 1st and 2nd images suited for multicoloured pocketted?

A5 **AGENT:** Yes

U5 **SHOPPER:** And for melange?

A6 **AGENT:** No

U6 **SHOPPER:** Show me more images of the 3rd product in some different directions

A7 **AGENT:** Sorry i couldnt find any other orientations for this image

U7 **SHOPPER:** I dont like the 4th image . Show me something in fit as in this image

Figure 1: Example chatlog depicting multi-modal user-agent interaction in a dialogue session from the MMD dataset. The system needs to ground knowledge to generate responses related to product-specific attributes. We focus on textual response generation given a fixed-size conversational history.

logue. In contrast, our system grounds on a KB while also conditioning on previous dialogue context which is multimodal in nature, consisting of both textual and visual communication between the user and the system. We formulate our KB input from a database query (triggered by the system) similar to Sha et al. (2018), as described in Section 3.2.

Our model belongs to the encoder-decoder paradigm where sequence-to-sequence models (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015) have become the de-facto standard for natural language generation. However, they tend to ignore the conversational history in a dialogue. The Hierarchical Recurrent Encoder Decoder (HRED) architecture (Serban et al., 2016, 2017; Lu et al., 2016) addresses this limitation by using a context recurrent neural network (RNN), forming a hierarchical encoder. We build upon these HRED models and refer to them as Text-only HREDs (T-HRED) in the following. Our model is most similar to the Multimodal HRED (M-HRED) of Saha et al. (2017), with context and KB extensions (see Section 3).

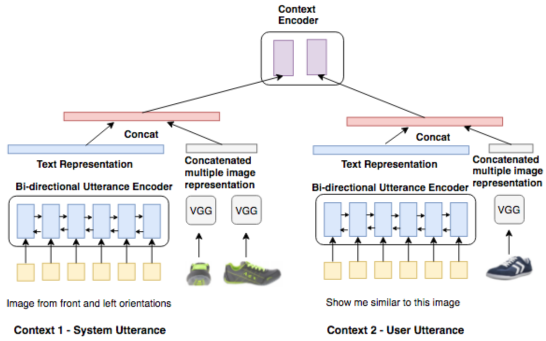

Figure 2: Schematic diagram of hierarchical encoder described in Section 3.1. Figure 3 depicts full pipeline of the model using knowledge base input. In contrast to Saha et al. (2017), we model over multiple images in a contextual dialogue turn by combining all 'local' representations of multiple images to a 'global' image representation per turn. We show a context of 2 turns for simplicity.

3 Knowledge grounded Multimodal Conversational model

While Saha et al. (2017) propose Multimodal HRED (M-HRED) by extending T-HRED to include visual context over images, they do not ground their dialogue context over an external database. Also, they limit the visual information by 'unrolling' multiple images to just use the last image of a single turn. For example in Figure 1, Saha et al. (2017) consider only the last image of trousers as visual context in Agent's response A4. In contrast, we include all the images in a single turn using a linear layer (see Agarwal et al. (2018) for a detailed analysis).

In addition, we devise a mechanism to ground our textual responses on a KB; Figure 3 depicts the full pipeline of our model. We combine textual and visual representations at the encoder level and pass it through the HRED's context encoder (cf. Figure 2), which learns the backbone of the conversation (see Section 3.1). Subsequently, we inject knowledge from the KB at the decoder level in each timestep (see Sections 3.2 and 3.3).

Formally, we model a dialogue as a sequence of utterances (turns) which are considered as sequences of words and images:

$$P_\theta(t_1, \dots t_N) = \prod_{n=1}^{N} P_\theta(t_n | t_{<n}) \qquad (1)$$

Here t_n represents the n-th utterance in a dialogue.

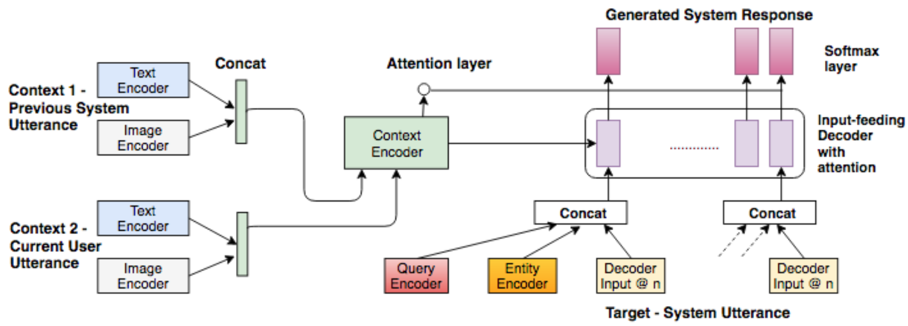

Figure 3: The full encoder-decoder pipeline of our model. While we have early fusion of textual and image representations (which act as input to the context encoder, see Figure 2), we employ late fusion of the knowledge base vector at the decoder level. For simplicity, we show a context of 2 turns.

The whole model is trained using cross entropy on next-word prediction:

$$J(\theta) = -\sum_{n=1}^{N} logP(y_n|y_0 \dots y_{t-1}) \qquad (2)$$

In the following, we explain all the different components of our model. We use the following notation: f_θ^{text},f_θ^{cxt}, f_θ^{query}, f_θ^{ent} and f_θ^{dec} are all GRU cells (Cho et al., 2014) and g_θ^{enc} is a Convolutional Neural Network (CNN) image encoder. θ represent our model weights. $w_{n,m}$ is the m-th word in the n-th textual utterance. Similarly, $q_{m,n}$ and $c_{m,n}$ represent input at each timestep in the query and entity encoder (see Section 3.2).

3.1 Hierarchical Encoder

The encoder is formed of the following modules:

Utterance (Text) encoder: We pass each utterance (previous system responses as well as current user query) in a given context through a text encoder. We use bidirectional GRU (f_θ^{text}) to generate the textual representation h_{n,M_n}^{text} (cf. Eq. (3)). These textual representations are combined with image representations in each turn, forming the input for the context encoder.

$$h_{n,m}^{text} = f_\theta^{text}(h_{n,m-1}^{text}, w_{m,n}); \quad h_{n,0}^{text} = 0 \quad (3)$$

Image encoder: We first extract the 'local' image representations for all images in a dialogue turn (denoted by $g_\theta^{enc}(img_k)$ in Eq. (4)) and concatenate them together.[1] This concatenated vector is passed through a linear layer to form the 'global' image context for a single turn, denoted by h_n^{img}.

$$h_n^{img} = l^{img}([g_\theta^{enc}(img_1), \dots g_\theta^{enc}(img_k)]) \quad (4)$$

Context encoder: The final hidden representations from both text encoder h_{n,M_n}^{text} and image encoder h_n^{img} are concatenated for each turn and serve as input to the context RNN (cf. Eq. (5)). On top of the text and image encoder, this builds a hierarchical encoder modelling the dialogue history. The final hidden state of the context RNN h_N^{cxt} acts as the initial state of the decoder RNN defined in Section 3.3.

$$h_n^{cxt} = f_\theta^{cxt}(h_{n-1}^{cxt}, [h_{n,M_n}^{text}, h_n^{img}]); \; h_0^{cxt} = 0 \quad (5)$$

3.2 Knowledge base (KB) input

The KB vector h_n^{kb} in Eq. (8) is formed by concatenating the h_n^{query} and h_n^{ent} representations. While our approach is modelled around the MMD dataset which provides contextual KB queries and profiles of celebrities endorsing specific products, it can be applied to other KBs with encoded queries and (optionally) properties of relevant entities.

$$h_n^{query} = f_\theta^{query}(h_{n-1}^{query}, q_{m,n}) \qquad (6)$$

$$h_n^{ent} = f_\theta^{ent}(h_{n-1}^{ent}, c_{m,n}) \qquad (7)$$

$$h_n^{kb} = [h_n^{query}, h_n^{ent}] \qquad (8)$$

$$h_0^{query} = 0; \quad h_0^{ent} = 0 \qquad (9)$$

Query encoder: Each chat session contains multiple queries to the database which retrieve the relevant product suited to user requirements at specific turn. We replicate this query for subsequent dialogue turns until a new query is triggered by the system. This query acts as knowledge base for the model at each turn. We show a sample input to the model in Figure 4. We used unidirectional GRU cell to encode the query input h_n^{query}.

Entity encoder: The input to the entity encoder is a list of entities relevant to the query at hand (see Figure 5). GRU cells are used to produce the resulting h_n^{ent}. Specifically, the MMD dataset categorises products into synonym sets (synsets) and

[1]We used the VGGnet (Simonyan and Zisserman, 2015) CNN to obtain the local image representations. Since the number of images in a turn is ≤ 5, we consider zero vectors in the absence of images.

Query:
```
"search_criteria": {
    "name": {"driving shoes": 1.0},
    "fit":{"tight": 1.0},
    "brand": {"cirohuner": 1.0},
    "image_type":{"front": 1.0},
    "gender": {"men": 1.0},
    "print": {"chain": 1.0}
}
```
Knowledge base input:
```
name driving shoes fit tight brand
cirohuner image_type front gender
men print chain
```
Figure 4: Sample query to the database and corresponding knowledge base input vector.

1.
```
User: what kind of trousers are
endorsed by celebrity cel_237?
Intent: celebrity
Subintent :does_celebrity_endorse_n
Celebrity: cel_237
Celebrity input: boxer briefs
```
2.
```
User: which of the celebrities
usually wear similar looking canvas
shoes as in the 2nd image
Intent: celebrity
Subintent: which_celebrity_endorses_n
Synset: canvas shoes
Celebrity input:
cel_987 cel_2 cel_316 cel_101
```
Figure 5: Two input scenarios for the entity encoder depending on the fine grained user intent. If there is no 'celebrity' intent, we have an empty string as input to the entity encoder.

provides a list of celebrities endorsing each synset (see Section 4.1 for details).

This input is used specifically for the 'celebrity' intent in our model, where the user asks about celebrities endorsing a product. For each target prediction with celebrity intent, we first extract the relevant celebrity profiles using basic pattern matching over the user utterance. For each of the celebrities in the user query, we order the corresponding synsets by their probability of endorsement. If no celebrity is found, we use synset information from the query to extract celebrities which endorse the corresponding synset.

3.3 Input feeding decoder

We use an input feeding decoder with the attention mechanism of Luong et al. (2015). We concatenate the KB input h_n^{kb} with the decoder input (cf. Eq. (10), where $h_{n,0}^{dec} = h_N^{cxt}$). The rationale behind this late fusion of KB representation is that KB input remains the same for a given context and does not change on each turn. On the other hand, images and textual response together form a context in a dialogue turn and thus we fuse them early at the encoder level. The decoder is trained using cross-entropy loss defined in Eq. (2).

$$h_{n,m}^{dec} = f_\theta^{dec}(h_{n,m-1}^{dec}, w_{n,m}, h_{n-1}^{cxt}, h_{n-1}^{kb}) \quad (10)$$

4 Experiments and Results

4.1 Dataset

Our work is based on the Multimodal Dialogue (MMD) dataset (Saha et al., 2017), which consists of 150k chat sessions.[2] User queries can be complex from the perspective of multimodal task-specific dialogue, such as "Show me more images of the 3rd product in some different directions". However, it also heavily relies on the external KB to answer product attributes related to user queries, such as "What is the brand/material of the suit in 3rd image?" or "Show something similar to 1st result but in a different material". This dataset contains raw chat logs as well as metadata information of the corresponding products. Around 400 anonymised celebrity profiles have been introduced in the system to emulate endorsement in recommendation, such as "What kind of slippers are endorsed by cel_145?". For each dialogue turn, there are manual annotations of the user intent available. We use the intents to construct celebrity encodings. On average, each session contains 40 dialogue turns. The system response depends on the intent state of the user query and on average contains 8 words and 4 images per utterance. We created our own version of the dataset from the raw chat logs of the dialogue session and metadata information. As discussed in Section 3.1, this was necessary to model the visual context over multiple images. We created the KB input to our model as described in Section 3.2 from the raw chat logs and the metadata information.

4.2 Implementation

We used PyTorch[3] (Paszke et al., 2017) for our experiments.[4] We did not use any kind of delexicalisation[5] and rely on our model to directly learn

[2]We used the same training-development-test split as provided by the dataset authors.

[3]https://pytorch.org/

[4]Code can be found at: https://github.com/shubhamagarwal92/mmd

[5]Replacing specific values with placeholders (Henderson et al., 2014).

from the conversational history and KB. All encoders and decoders are based on 1-layer GRU cells (Cho et al., 2014) with 512 as the hidden state size. We used the 4096 dimensional FC6 layer image representations from VGG-19 (Simonyan and Zisserman, 2015) provided by Saha et al. (2017). Adam (Kingma and Ba, 2015) was chosen as the optimizer, and we clipped gradients greater than 5. We experimented with different learning rates and settled on the value of 0.0004. Dropout of 0.3 is applied to all the RNN cells to avoid overfitting, and we perform early stopping by tracking the validation loss (with single trial for each experiment).

4.3 Analysis and Results

We evaluate our response generation using the BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007) and ROUGE-L (Lin and Och, 2004) automatic metrics.[6] We reproduce the baseline results from Saha et al. (2017) using their code and data-generation scripts.[7]

Model	Cxt	BLEU-4	METEOR	ROUGE-L
Saha et al. M-HRED*	2	0.3767	0.2847	0.6235
T-HRED	2	0.4292	0.3269	0.6692
M-HRED	2	0.4308	0.3288	0.6700
T-HRED–attn	2	0.4331	0.3298	0.6710
M-HRED–attn	2	0.4345	0.3315	0.6712
T-HRED–attn	5	0.4442	0.3374	0.6797
M-HRED–attn	5	0.4451	0.3371	0.6799
M-HRED–kb	2	0.4573	0.3436	0.6872
T-HRED–attn–kb	2	0.4601	0.3456	0.6909
M-HRED–attn–kb	2	0.4624	0.3476	0.6917
T-HRED–attn–kb	5	0.4612	0.3461	0.6913
M-HRED–attn–kb	5	**0.4634**	**0.3480**	**0.6923**

Table 1: Automated evaluation based on BLEU-4, METEOR and ROUGE-L metrics. Here, 'M' represents multimodality while 'T' stands for text-only model. 'attn' denotes use of attention and 'kb' signifies incorporating Knowledge Base input. 'Cxt' represents context size for the dialogue history.
*Saha et al. was trained on a different version of the dataset, as discussed in Section 3.

Table 1 summarises the results for our M-HRED model without incorporating KB information. Attention-based models consistently outperform their counterparts. Adding the visual inputs does not lead to major improvements (M-HRED vs. T-HRED for a given context). However, grounding in KB gave a stark uplift (M-HRED–attn–kb vs. M-HRED–attn) for a given context size. Adding KB input boosts performance more for a shorter context compared to longer context. It can be conjectured that the longer context contains some of the information that is in the KB queries and so there is less impact of the KB input when we include the longer context. Compare the difference for M-HRED–attn–kb vs. M-HRED–attn for a context of 2 (3 BLEU points) vs. 5 (2 BLEU points) in Table 1. Conversely, longer context improves more the models without KB queries.

Intent	Model	BLEU-4
show-similar-to	M-HRED–attn	0.9998
	M-HRED–attn–kb	1.0
sort-results	M-HRED–attn	0.9188
	M-HRED–attn–kb	0.9384
suited-for	M-HRED–attn	0.6151
	M-HRED–attn–kb	0.6216
show-orientation	M-HRED–attn	0.5388
	M-HRED–attn–kb	0.5854
buy	M-HRED–attn	0.2665
	M-HRED–attn–kb	0.3179
ask-attribute	M-HRED–attn	0.4960
	M-HRED–attn–kb	0.5934
celebrity	M-HRED–attn	0.2671
	M-HRED–attn–kb	0.2725

Table 2: BLEU scores for the entire corpus predictions for specific intents with a context of 5.

In summary, our best performing model (M-HRED–attn–kb) outperforms the model of Saha et al. (2017) by 9 BLEU points. We also analysed our generated outputs for different user intents, as shown in Table 2. As assumed, intents such as 'show-similar-to' and 'sort-results' are relatively easy from the perspective of NLG, requiring no information about the product description; our model matches the reference almost perfectly.

We found great improvements for the 'ask-attribute' intent where the KB-grounded model could answer correctly questions related to brand or colour and other attributes of the product, which resulted in an increase of 10 BLEU points on test instances with this user intent (M-HRED–attn–kb compared to M-HRED–attn). Similarly, in the example related to the 'buy' intent in Table 3, our model is able to learn that the product bought by the user is 'kurta', which probably cannot be captured by the visual features. Hence, M-HRED–attn produces 'jeans' on the output. M-HRED–attn–kb on the other hand learns this information from the KB. We also found that our BLEU score for the 'show-orientation' intent has decreased w.r.t. to the non-KB-grounded model. A detailed

[6] We used the evaluation scripts provided by (Sharma et al., 2017).

[7] https://github.com/amritasaha1812/MMD_Code

Intent	Model	Example Text
show-similar-to		**Text context**: yes. | show me something similar to the 1st image but in a different material
		Gold Target: the similar looking ones are
		KB: *name[casual-trousers] gender[women] brand[antigravity] synsets[casual-trousers]*
	M-HRED–attn	**Predicted:** the similar looking ones are
	M-HRED–attn–kb	**Predicted:** the similar looking ones are
sort-results		**Text context**: sorry i dont seem to have anything in loop but would you like in slim | sort these by best seller ranking
		Gold Target: sorry i do not have the best seller ranking information for these
		KB: *name[casual-trousers] gender[women] brand[antigravity] synsets[casual-trousers]*
	M-HRED–attn	**Predicted:** sorry i dont have the best seller ranking information for the rest
	M-HRED–attn–kb	**Predicted:** sorry i do not have the best seller ranking information for these
suited-for		**Text context**: no. | which care might go well with this dupatta?
		Gold Target: it will suit well for professional and handwash care
		KB: *name[dupatta] gender[women] synsets[dupatta] image-type[front] care[handwash]*
	M-HRED–attn	**Predicted:** it will suit well for gentle hand care
	M-HRED–attn–kb	**Predicted:** it will suit well for *hand wash care*
show-orientation		**Text context**: | show me more images of the 3rd product from some different orientations
		Gold Target: sorry i couldnt find any other orientations for this image
		KB: *name[casual-trousers] gender[women] brand[antigravity] synsets[casual-trousers]*
	M-HRED–attn	**Predicted:** image from the front, right, back and left orientations respectively
	M-HRED–attn–kb	**Predicted:** sorry i couldnt find any other orientations for this image
buy		**Text context**: | i like the 4th one. i think ill buy that one
		Gold Target: absolutely , thats a great kurta
		KB: *name [kurta] color [green] gender [men] synsets [kurta] image-type [front]*
	M-HRED–attn	**Predicted:** absolutely , i think thats a great jeans
	M-HRED–attn–kb	**Predicted:** absolutely , i think thats a great *kurta*
ask-attribute		**Text context**: yes. | what is the brand in the 1st result?
		Gold Target: the blouse in the 1st image has alfani brand
		KB: *name [blouse] brand [alfani] synsets [blouse] image-type [look] gender [women]*
	M-HRED–attn	**Predicted:** the brand in 1st image is topshop
	M-HRED–attn–kb	**Predicted:** the brand in 1st image is *alfani*
celebrity		**Text context**: yes. celebrities cel_779, cel_10 and cel_513 also endorse this type of cufflinks | and celebrity cel_603 for the 1st?
		Gold Target: yes
		KB Query: *name[casual-trousers] gender[women] synsets[casual-trousers]*
		KB Entity: *scarf earrings casual trousers casual shirt*
	M-HRED–attn	**Predicted:** no.
	M-HRED–attn–kb	**Predicted:** yes.

Table 3: Examples of predictions corresponding to different user intents, showcasing the effect of grounding in KB. We show textual context as well as relevant knowledge base input (and omit image context) for brevity's sake. While our model uses a context of 5, for simplicity, we show only 2 previous turns.

probe found that the orientations for retrieved images may not directly follow the description in the query (KB). There are other intents for which even KB does not help, such as those requiring user modelling.

5 Conclusion and Future Work

This work focuses on the task of textual response generation in multimodal task-oriented dialogue system. We used the recently released Multimodal Dialogue (MMD) dataset (Saha et al., 2017) for experiments and introduced a novel conversational model grounded in language, vision and Knowledge Base (KB). Our best performing model outperforms the baseline model (Saha et al., 2017) by 9 BLEU points, improving context modelling in multimodal dialogue generation. Even though our model outputs showed a substantial improvement (over 3 BLEU points) on incorporating KB information, integrating visual context still remains a bottleneck, as also observed by Agrawal et al. (2016); Qian et al. (2018). This suggests the need for a better mechanism to encode visual context.

Since our KB-grounded model assumes user intent annotation and KB queries as additional inputs, we plan to build a model to provide them automatically.

Acknowledgments

This research received funding from Adeptmind Inc., Montreal, Canada and the MaDrIgAL EPSRC project (EP/N017536/1). The Titan Xp used for this work was donated by the NVIDIA Corp.

References

Shubham Agarwal, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. 2018. Improving context modelling in multimodal dialogue generation. In *Proc. INLG*. To appear.

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proc. EMNLP*, pages 1955–1960.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proc. ICCV*, pages 2425–2433.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proc. ICLR*.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *CoRR abs/1605.07683*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proc. EMNLP*, pages 1724–1734.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proc. ICCV*, volume 2.

Hao Fang, Hao Cheng, Elizabeth Clark, Ariel Holtzman, Maarten Sap, Mari Ostendorf, Yejin Choi, and Noah A Smith. 2017. Sounding Board–University of Washington's Alexa Prize submission. In *Alexa Prize Proceedings*.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2017. A knowledge-grounded neural conversation model. *CoRR abs/1702.01932*.

M. Henderson, B. Thomson, and S. Young. 2014. Robust dialog state tracking using delexicalised recurrent neural networks and unsupervised adaptation. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 360–365.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, Lawrence C. Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. *Proc. NAACL*, pages 1233–1239.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proc. ICCV*, pages 3128–3137.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. ICLR*.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proc. 2nd Workshop on Statistical Machine Translation*, pages 228–231.

Oliver Lemon, Kallirroi Georgila, James Henderson, and Matthew Stuttle. 2006. An ISU dialogue system exhibiting reinforcement learning of dialogue policies: generic slot-filling in the talk in-car system. In *Proc. EACL*, pages 119–122.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proc. ACL*, page 605.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Proc. NIPS*, pages 289–297.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *Proc. EMNLP*, pages 1412–1421.

Ioannis Papaioannou, Amanda Cercas Curry, Jose L Part, Igor Shalyminov, Xinnuo Xu, Yanchao Yu, Ondrej Dušek, Verena Rieser, and Oliver Lemon. 2017. Alana: Social dialogue using an ensemble model and a ranker trained on user feedback. *Alexa Prize Proceedings*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pages 311–318.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS-W*.

Xin Qian, Ziyi Zhong, and Jieli Zhou. 2018. Multimodal machine translation with reinforcement learning. *CoRR abs/1805.02356*.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. 2018. Conversational AI: The science behind the Alexa Prize. *CoRR abs/1801.03604*.

Verena Rieser and Oliver Lemon. 2010. Natural language generation as planning under uncertainty for spoken dialogue systems. In *Empirical methods in natural language generation*, pages 105–120. Springer.

Verena Rieser and Oliver Lemon. 2011. *Reinforcement learning for adaptive dialogue systems: a data-driven methodology for dialogue management and natural language generation*. Springer.

Amrita Saha, Mitesh Khapra, and Karthik Sankaranarayanan. 2017. Multimodal dialogs (MMD): A large-scale dataset for studying multimodal domain-aware conversations. *CoRR abs/1704.00200*.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proc. AAAI*, pages 3776–3783.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proc. AAAI*, pages 3295–3301.

Lei Sha, Lili Mou, Tianyu Liu, Pascal Poupart, Sujian Li, Baobao Chang, and Zhifang Sui. 2018. Order-planning neural text generation from structured data. *Proc. AAAI*.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR abs/1706.09799*.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *Proc. ICLR*.

Satinder P Singh, Michael J Kearns, Diane J Litman, and Marilyn A Walker. 2000. Reinforcement learning for spoken dialogue systems. In *Proc. NIPS*, pages 956–962.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proc. NIPS*, pages 3104–3112.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. MovieQA: Understanding stories in movies through question-answering. In *Proc. CVPR*, pages 4631–4640.

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015a. Sequence to sequence-video to text. In *Proc. ICCV*, pages 4534–4542.

Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2015b. Translating videos to natural language using deep recurrent neural networks. *Proc. NAACL*, pages 1494–1504.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proc. CVPR*, pages 3156–3164.

Zhuoran Wang and Oliver Lemon. 2013. A simple and generic belief tracking mechanism for the dialog state tracking challenge: On the believability of observed information. In *Proc. SIGDIAL*, pages 423–432.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proc. EACL*, pages 438–449.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. ICML*, pages 2048–2057.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.