# A KNOWLEDGE-GROUNDED MULTIMODAL SEARCH-BASED CONVERSATIONAL AGENT

Shubham Agarwal * Ondřej Dušek Ioannis Konstas Verena Rieser
Interaction Lab, Department of Computer Science, Heriot-Watt University, Edinburgh, UK
* Adeptmind Scholar, Adeptmind Inc., Toronto, Canada

EMNLP 2018

## OVERVIEW

**Task:** Multimodal search-based dialogue system

- Task oriented dialogue system in e-commerce setting
- Based on recently released **MultiModal Dialogue (MMD)** dataset
- **Multimodal HRED with attention** for textual response generation grounded in vision, language and knowledge base input

## DATASET

- Raw chatlog of an user-agent interaction in the fashion domain (150k chat sessions with 40 dialogue turns per session)

**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?

## KB GROUNDED MULTIMODAL CONVERSATIONAL MODEL

- Hierarchical Encoder
  - Utterance (Text) encoder
  - Image encoder
  - Context encoder

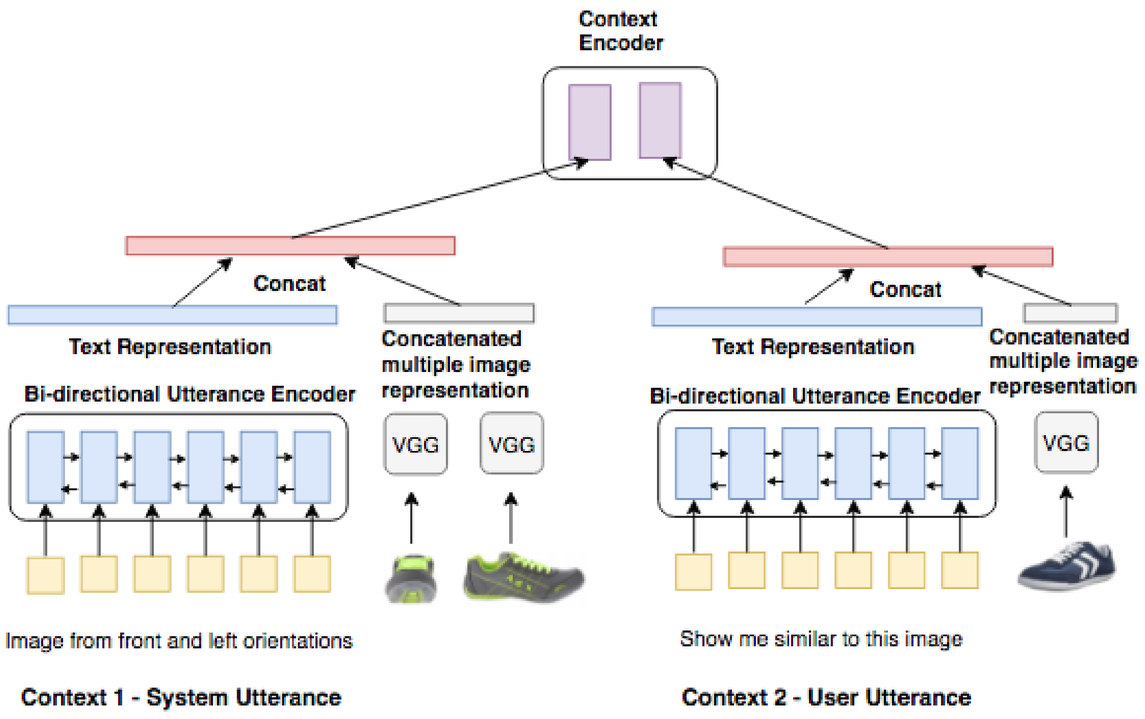


**Fig 1: Early fusion of textual and visual representations at the encoder level. Learns the backbone of the conversation.**

- Knowledge base (KB) encoder
  - Query encoder
  - Entity encoder
- Input feeding decoder

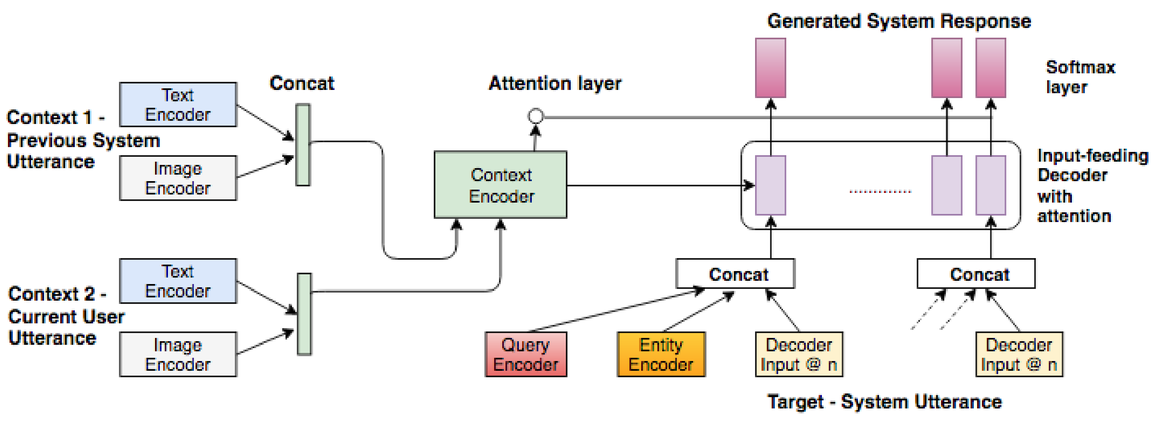


**Fig 2: Full encoder-decoder pipeline of model. Late fusion of the knowledge base vector at the decoder level.**

## EVALUATION AND RESULTS

| 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: Automatic evaluation based on BLEU-4, METEOR and ROUGE-L**

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

**Table 2: BLEU scores for the entire corpus predictions for specific intents**

## PREDICTIONS FOR DIFFERENT INTENTS

| Intent | Model | Example Text |
|---|---|---|
| **show-similar-to** | | **Text context**: yes. \| show me something similar to the 1st image but in a different material<br>**Gold Target:** the similar looking ones are<br>**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<br>**Gold Target:** sorry i do not have the best seller ranking information for these<br>**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?<br>**Gold Target:** it will suit well for professional and handwash care<br>**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<br>**Gold Target:** sorry i couldnt find any other orientations for this image<br>**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<br>**Gold Target:** absolutely , thats a great kurta<br>**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?<br>**Gold Target:** the blouse in the 1st image has alfani brand<br>**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?<br>**Gold Target:** yes<br>**KB Query:** *name[casual-trousers] gender[women] synsets[casual-trousers]*<br>**KB Entity:** *scarf earrings casual trousers casual shirt* |
| | M-HRED–attn | **Predicted:** no. |
| | M-HRED–attn–kb | **Predicted:** yes. |

## CONCLUSION

- A novel conversational model grounded in language, vision and KB
- Best performing model (M-HRED–attn–kb) outperforms baseline by 9 BLEU points. Grounding in KB gave stark uplift (M-HRED–attn–kb vs. M-HRED–attn)
- Integrating visual context still remains a bottleneck. Need better visual models to encode the image representations for multiple similar-looking images
- Code available at `https://github.com/shubhamagarwal92/mmd`