DSTC8 Track Proposal: Dialog State Tracking for Conversational Image Editing

Trung Bui, Seokhwan Kim, Franck Dernoncourt, Quan Tran

Adobe Research, San Jose, California

1 Background and Motivation

We propose a new challenge for the dialog research community: dialog state tracking for conversational image editing. Conversational image editing [1] is a novel application domain that combines language and vision as well as action execution in task-oriented scenarios. The dialog application goes beyond traditional slot-filling systems for restaurant search or movie booking. In addition, it is also more challenging compared to visual dialog [2], since in conversational image editing the image content is being changed as the conversation unfolds.

2 Proposed Task

The main task of this challenge is to track dialog states in the form of frame structures as defined in [3]. For each turn in a given dialog session, the tracker should fill a frame considering the following details:

1. Slot-value pairs to represent the detailed user intentions
2. Whether the current turn should be associated with a new frame or not

Table 2 in Appendix shows examples of dialog annotations for this task.

2.1 Data

Our proposed task will be based on DialEdit [4] dataset which includes spoken dialogues collected from Skype calls between users (who request image edits) and wizards (who perform the edits). In each session, a wizard performed an image editing task based on the user requests from the conversation. They shared the screen of an image editing software running on the wizard's machine. A total of 28 users and 2 wizards participated in the data collection (Table 1). All the recorded dialogues have been manually transcribed and annotated with both SLU and frame structures.

For this challenge, the DialEdit corpus will be divided into three parts: training, development, and test sets. Both training and development sets will be released at the beginning of the challenge period with all the utterances and the labels, while only the raw utterances will be available for the test set during the evaluation phase. In addition to the corpus, an ontology and a handbook will be provided to help participants to understand the task details.

# users	28
# dialogues	129
# user utterances	8,890
# Wizard utterances	4,795
# time (raw, in minutes)	858
# user tokens	59,653
# user unique tokens	2,299
# Wizard tokens	26,284
# Wizard unique tokens	1,310
# total unique tokens	2,650

Table 1: Statistics of DialEdit dataset

2.2 Evaluation

A system should generate the frame tracking output for every user utterance in a given dialogue session. Only the utterances from the beginning of the session to the current turn can be used to predict the frame values. Any information from the future turns is NOT allowed to be considered at a given turn. The system outputs will be evaluated by the following same metrics as DSTC4 [5]:

- Accuracy: Fraction of turns in which the tracker's output is equivalent to the gold standard labels
- Precision/Recall/F-measure
  - Precision: Fraction of slot-value pairs in the tracker's outputs that are correctly filled
  - Recall: Fraction of slot-value pairs in the gold standard labels that are correctly filled
  - F-measure: The harmonic mean of precision and recall

2.3 Other Resources

We will provide the following scripts and tools for the challenge participants:

- Baseline tracker: a simple neural network implementation
- Evaluation scripts: validating the formats and calculating the scores
- Data loaders: getting the information from the corpus and the ontology

All the resources including the corpus and the scripts will be released on our Github repository.

3 Organizers

- Trung Bui, Adobe Research, bui@adobe.com
- Seokhwan Kim, Adobe Research, seokim@adobe.com
- Franck Dernoncourt, Adobe Research, dernonco@adobe.com
- Quan Tran, Adobe Research, qtran@adobe.com

4 Appendix

A simplified example of the conversation between the user and the wizard is described in Table 2.

Speaker	Utterance	SLU annotations	Frame annotations
User	Brighten the yellow dog	**Dialog act:** IER_N **Intent:** adjust **IOB:** [action: brighten] [object 1: the yellow dog].	**Frame ID:** 1, **Intent**:adjust **Attribute:** brightness, **Object:** 1
Wizard	Okay. I've done.		
User	Move it to the left	**Dialog act:** IER_N **Intent:** move **IOB:** [action: move] [object 1: it] [location 1: to the left].	**Frame ID:** 2, **Intent**: move **Object:** 1, **Location:** 1
Wizard	Are you okay with this?		
User	A bit more to the left	**Dialog act:** IER_U, **Intent:** move **IOB:** [value: a bit more] [location 1: to the left].	**Frame ID:** 2, **Intent**: move **Object:** 1, **Location:** 1, **Value:** +
Wizard	Ok. Anything else?		
User	darken the building on the right	**Dialog act:** IER_N, **Intent:** adjust **IOB:** [action: darken] [object 2: the building on the right].	**Frame ID:** 3, **Intent**: adjust **Attribute:** brightness, **Object:** 2, **Value:** -

Table 2: An example dialog with SLU and frame annotations

References

[1] R. Manuvinakurike, T. Bui, W. Chang, and K. Georgila, "Conversational Image Editing: Incremental Intent Identication in a New Dialogue Task," in *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, (Melbourne, Australia), pp. 284–295, Association for Computational Linguistics, July 2018.

[2] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra, "Visual Dialog," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] L. El Asri, H. Schulz, S. Sharma, J. Zumer, J. Harris, E. Fine, R. Mehrotra, and K. Suleman, "Frames: a corpus for adding memory to goal-oriented dialogue systems," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 207–219, Association for Computational Linguistics, 2017.

[4] R. Manuvinakurike, J. Brixey, T. Bui, W. Chang, R. Artstein, and K. Georgila, "DialEdit: Annotations for Spoken Conversational Image Editing," in *Proceedings of the 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, (Santa Fe, New Mexico), Association for Computational Linguistics, Aug. 2018.

[5] S. Kim, L. F. D'Haro, R. E. Banchs, J. D. Williams, and M. Henderson, "The fourth dialog state tracking challenge," in *Dialogues with Social Robots*, pp. 435–449, Springer, 2017.