RESPONSE TO REVIEWERS

"Entity Resolution in Situated Dialog with Unimodal and Multimodal Transformers"

Submitted to Transactions on Audio, Speech and Language Processing

***Alejandro Santorum Varela**, from Cambridge University Engineering Dept. (Cambridge, UK)*
***Svetlana Stoyanchev**, from Toshiba Europe Ltd. (Cambridge, UK)*
***Rama Doddipatla**, from Toshiba Europe Ltd. (Cambridge, UK)*
***Simon Keizer**, from Toshiba Europe Ltd. (Cambridge, UK)*
***Kate Knill**, from Cambridge University Engineering Dept. (Cambridge, UK)*

We would like to thank the reviewers and the AE for their insightful comments and suggestions. The submitted paper has been revised in order to address all the previous imperfections. We hope that the updated version is clearer, more concise and helpful for all the audio, speech and language processing community.

1. Reviewers comments

Reviewer: 1

"A couple of points that could be clarified to enhance the clarity of your results. In the analysis of the Mentioned vs. New references (in Section V.C)"

1. *"Firstly, it would be beneficial to provide additional details regarding the performance of your auxiliary task. Specifically, it would be helpful to know the prediction accuracy or F1 scores of the auxiliary task, as this would provide context to the observed benefits for the main task. Without this information, it is difficult to assess the relative impact of the auxiliary task on the performance of the Mentioned vs. New references tasks."*

> The prediction accuracy of the auxiliary task head has been added in Section V.C: Mentioned vs. new references. The prediction accuracy is 99.1% for the UNITER-based model and 98.1% for the BART-based model.

2. *“Secondly, it would be helpful to provide more information on the dataset statistics that are relevant to your auxiliary task (i.e. the number of objects referred to by the user, etc.) Providing these statistics would give the reader a better understanding of the task and the performance of the auxiliary task.”*

> In Section V.C (Mentioned vs. new references), we have included the following two statistics: the proportion of targets and the proportion of new/mentioned object references in the SIMMC2.0 dataset. As we point out, 50.4% of the examples in this set have no targets (i.e., no referred items by the user in their last turn), 23.1% of the examples have a single target, *i.e.*, the user refers to one object in their last utterance, and in the rest (26.5% of the examples) the user refers to two objects in their last turn. Additionally, we highlight that about 58.5% of the target objects (ground-truth objects referred to by the user) have been previously mentioned in the dialog. These cases can be resolved using dialog context. For the other 41.5% target objects there is no information within the conversation context (they are “new”), so these cases require understanding of the scene.

Reviewer: 2

1. *“(Major) The proposed method does not seem novel (see comment on "novelty"). The paper mentions that the approach is applicable in any situated AI application, but only evaluated the method on one dataset and one of the four subtasks. Further experiments are needed to show generalizability of the method.” Additional comment in “novelty” question: “The paper first reviewed two established models (UNITER, BART) for the coreference prediction task on the SIMMC2 dataset, then proposed adding an auxiliary task that predicts the number of objects mentioned in an utterance. Even though the modified models result in higher F1 score, the modification was incremental and very specific to this one dataset and one of the sub-tasks.”*

> First, we have modified the Abstract, the introduction (Section I) and the conclusion (Section VII) to emphasize that our work is focused on analyzing two different approaches to the reference resolution task. We use as a baseline two top performing models that participated in the DSTC10 competition and analyze them in-depth by running the experiments in several scenarios to study the generalizability of the models. In particular, we compare the performance of the models in in-domain scenarios (seen objects and scenes at training when inferring), in-domain-held-out scenarios (seen objects but unseen scenes) and in out-of-domain setting, where the objects and the scenes at inference haven´t been seen at training time. This set of experiments of Sections V.A and V.B allow us to study the generalization capacity of the different models and how to improve them in unknown settings. In addition, we propose the modifications that improve their performance over baselines.

2. *“(Minor) I noticed the description of the proposed method is not mentioned until much later in the paper. This makes it difficult for the readers to fully grasp the contribution early on.”*

> To address this issue, we have added details in the introduction highlighting the main experiments, describing the proposed modifications and underlining the main contributions of the work. In addition, we have improved the outline (last paragraph of the introduction),

explaining how the paper is divided, so the future reader can visit the sections of particular interest.

3. *“(Minor) Please consider adding margin of errors in the evaluation results.”*

> We have included the margin of errors (standard errors) in the majority of the evaluation result tables. In particular, in Table VI, Table VII, Table VIII and Table IX we show the standard error.

2. Additional AE comments

“While there is a concern over the novelty of the work raised by one of the reviewers, the empirical evaluation is thorough which provides important insights to the field, as pointed by another reviewer. The recommendation is therefore to include discussions on the general applicability of the work beyond the demonstrated work, potentially on a different dataset (if applicable) or various experimental setups within the current dataset.”

> As described before when answering *Reviewers 2 comment no. 1*, Sections V.A and V.B include experiments to analyze the considered and improved models under different scenarios (in-domain, in-domain with unseen scenes and out-of-domain). These experiments aim to study and analyze the general applicability of the models in unknown settings by modifying the standard splitting of the considered SIMMC2 dataset in the DSTC10. Our experiments show that the unimodal approach outperforms the multimodal one on in-domain experiments. However, this method is only effective when objects are seen in training. In contrast, the multimodal approach outperformed the unimodal one in a cross-domain setting, showing potential for generalization.