	Full	Uncorrelated	Success Only
Random	14.28	14.28	14.28
Context Only (MLP)	27.90 ± 0.6	28.74	29.59
Context Only (RN)	31.94 ± 0.9	30.22	32.40
Context + Dialogue (MLP)	40.27 ± 1.3	40.89	43.82
Context + Dialogue (RN)	43.09 ± 0.8	44.00	49.44
Humans	-	82.50	90.79

Table 4: Results of the target selection experiment. Models are trained 10 times initialized with different seeds for the Full testset, and the models with best validation loss are used for the additional testset results (Uncorrelated and Success Only).

Following common practice, we split the dataset into training, validation and test set with a proportion of 8:1:1, and all models are tuned on the validation set. The loss function is calculated using cross entropy. All components of the neural networks consist of single layer with 128 hidden units, and dropout rate of 0.5 is applied at each layer to avoid overfitting. All parameters are initialized uniformly within the range of (-0.01, 0.01). Models are trained with the Adam optimizer (**?**) with initial learning rate of 0.001, and we clip gradients whose L^2 norm is greater than 0.1. The experiment is run 10 times initialized with different seeds, and we report the mean and standard deviation of the selection accuracies on the full testset.

For further analyses, models with the best validation loss in the previous experiment are also tested on two variants of the testset. First, we create an uncorrelated testset by randomly removing one from each correlated pair in the current testset (same dialogue but different context). Secondly, we further removed dialogues where players failed to coordinate on the same entity from the uncorrelated testset, since this may affect target selection performance. The statistical significance of the results for each pair of methods are tested on the uncorrelated testset using paired student's t-test. Finally, we take 100 random samples from the uncorrelated testset (including 76 successful) to report human performance based on average accuracy of two annotators.

5.3 Results

We show the results of our experiment in Table 4. As we can see, models trained only with the context embeddings perform significantly better than random (p-value $< 10^{-7}$). This verifies that we can indeed take advantage of selection bias to make better predictions.

In addition, we found that embedding context with Relation Network consistently outperforms MLP, but not at a statistically significant level (p-value > 0.1). Therefore, the simplest strategy of using MLP works decently, but a better architecture may improve the overall performance.

Finally, models trained with both context and dialogue embeddings significantly outperform models trained only with the context embeddings (p-value $< 10^{-9}$). This indicates that even our simplest models can learn to ground linguistic meanings based on the context to make better predictions. When the testset only includes successful cases, models perform better but human performance improves even more achieving over 90% accuracy. Overall, our target selection task is challenging due to the complexity of common grounding, and we still have a huge room for improvement.

6 Conclusion and Future Work

The main contributions can be summarized as follows:

- We proposed a simple and general idea of incorporating continuous and partially-observable context to the dialogue tasks, which makes common grounding difficult in a natural way.
- Following this idea, we formulated a novel dialogue task based on collaborative referring which enables clear evaluation and analysis of complex models.
- We collected a largescale dataset of 6,760 dialogues, which fulfills essential requirements of natural language corpora and will be publicly available online.
- Our analysis of the dataset verified the difficulty of common grounding and revealed various phenomena that need to be considered.
- We evaluated and analyzed simple baseline models on an important subtask of collaborative referring and showed that there is still room for further improvement.

In future work, we will evaluate and analyze dialogue models based on our task, especially to identify the current limitations of end-to-end approaches in terms of common grounding. Models can be trained in a variety of ways, including supervised learning, reinforcement learning with humans, and reinforcement learning based on *self-play* (**?**). Overall, we expect our task to be a fundamental testbed for developing dialogue systems with sophisticated common grounding abilities.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 16K12546,18H03297.

Nihil libero saepe molestias voluptatum consequatur blanditiis, molestiae nostrum sed quidem accusantium, soluta dolores tenetur sunt sequi similique reiciendis, dolorum repudiandae non consequatur labore porro laudantium, ratione ea cupiditate alias.Nulla voluptas ipsa placeat necessitatibus, sint fuga facilis a recusandae rerum officiis, voluptatem quas culpa a cum facilis quos omnis placeat atque?Animi laboriosam libero suscipit impedit laborum ipsam nostrum blanditiis facilis, expedita eligendi voluptatem tenetur cumque tempora voluptates quo reprehenderit, eligendi reiciendis neque ad magni veritatis explicabo optio inventore corrupti.Qui fugit sequi laudantium

maiores sint aspernatur voluptate natus porro in ea, eaque illum molestiae hic totam laboriosam adipisci alias, architecto