Modeling Common Ground in Color Reference Games with a Differentiable Neural Computer

Benjamin Newman
blnewman@stanford.edu

Julia Gong
jxgong@stanford.edu

Abstract

While most natural language models in recent years have been focused on performing specific contextual tasks, it's also of interest to find ways to incorporate prior knowledge and non-contextual, general world knowledge into speaker-listener interactions. In this paper, we present a framework for modeling common ground between speakers and listeners in the context of reference games. We present a Speaker-Listener model with separate networks for each of the agents performing a color reference game task, and we augment this simpler model with a differential neural computer (DNC) to model the common ground between the speaker and listener. The hypothesis was that initializing the DNC with non-contextual color encodings would allow the DNC model to achieve higher accuracies in the reference game task. Though this did not turn out to be the case, likely due to insufficient training and possibly unsuitable color representations, we hope that this work lays down the framework for future work in modeling common ground and using ungrounded knowledge to enhance performance in grounded language tasks.

1 Introduction

Whether writing a blog post, negotiating a car price, or having a conversation with friends, language provides a means to express desires and achieve goals. Despite this fact, much of the current research in natural language processing and understanding is focused on solving discrete tasks such as sentiment analysis or question answering. Such a narrow focus does not easily allow for consideration of the motivations behind why some one expresses a certain sentiment or why they are asking a question that needs answering. As artificial agents that use natural language are increasingly interacting with non-experts who have non-academic goals and intentions, these agents need to develop more robust models of human cognition, social structures, and language to effectively communicate with and understand the goals of their users.

Models that do explicitly take into account human goals do exist, and are ubiquitous in the cognitive science literature. One such model that has received a lot of attention recently is the Rational Speech Acts model (RSA) (Goodman and Frank, 2016). The RSA model codifies the idea that when communicating, one ought to consider the goals of an conversational partner by relying on the strong inductive bias that all reasoning related to communication is recursive. In other words, when a speaker is communicating with a listener, the speaker chooses an utterance from a set of possible utterances by considering which utterance will best make the listener understand what the speaker is thinking. RSA-style models have mostly been used in the context of reference games. In these scenarios, there are usually two agents and a number of items. One agent, the speaker, has to generate a natural language utterance to get the other agent, the listener, to select some item that only the speaker knows is the target. In addition to predicting outcomes in reference games, RSA models been found to predict human actions in a variety of scenarios ranging from hyperbole and metaphor to negotiation tactics (Lewis et al., 2017) (Kao et al., 2014). Recently, they have also been integrated with neural models and improved predictions in more realistic settings, such as with images and natural language utterances [(Cohn-Gordon et al.), (Monroe et al., 2017), (Vedantam et al.), (Andreas and Klein, 2016)].

While these models are able to capture human-like reasoning in these situations, they do not currently provide any way to incorporate memories from previous interactions between agents in the

model. Additionally, these models only capture language used in these grounded, reference game contexts, while humans tend to possess knowledge about more general abstract forms apart from these limiting contexts. In this project, our goals are two-fold. First, we seek to expand the recursive reasoning in the RSA model to include a form of memory by incorporating a differentiable neural computer (DNC) (Graves et al., 2016). Second, we plan to seed the DNC with an abstract vocabulary derived from an ungrounded context to try to improve performance at generating effective, human-like referring expressions in the context of a color reference game.

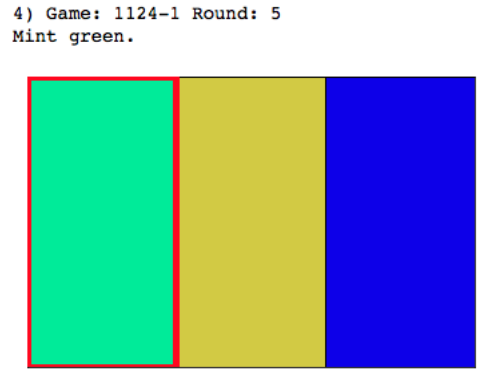

Figure 1: Sample game from (Monroe et al., 2017). The target color is boxed in red on the left.

2 Related Work

2.1 Common Ground Modeling

The field of common ground modeling is often intertwined with the field of knowledge modeling in general. Traditional approaches involve creating databases that are queried based on some salient quality of the knowledge that is being recorded. These approaches work when there are fixed, non-deterministic key value pairs or relationships, but in traditional knowledge modeling systems, these cannot be learned and therefore provide limited insight into how knowledge modeling actually works in the human brain (Devedzic, 2001).

2.2 Computational Pragmatics

Much of the related work comes from the field of computational pragmatics. As mentioned earlier, RSA has been successful in describing certain phenomena such as metaphor, hyperbole, negotiation, and reference game scenarios. There is some work extending RSA to multiple turn dialogues, where the speaker and listener have a conversation (Khani et al., 2018). This idea is most similar to ours because it involves longer-term planning, but there is no work that we know of so far that explicitly reasons across multiple games or rounds the way we are modeling with the differentiable neural computer.

2.3 Differentiable Neural Computer

The differentiable neural computer (DNC) model has provided a good way to incorporate long term memory into neural networks. Traditional recurrent neural networks such as LSTMs do have capacity for some long term memory, but their memory is weak over longer time scales (Khandelwal et al., 2018). DNCs address this problem and have been used for a range of tasks that traditional LSTMs have difficulty performing such as pattern copying, nearest neighbors, long-term planning, and question answering (Graves et al., 2016).

3 Methods

In this section, we describe the sources of our data and the models we propose.

3.1 Data

We are going to be addressing this problem in the context of a color reference game created by (Monroe et al., 2017). The data were collected in a human experiment performed as follows. Two people were shown three colored squares in a randomized order where the colors were of varying similarity. One participant (the speaker) had to generate a natural language utterance to try to get another participant (the listener) to select a target color that only the speaker knew. The data set contains approximately 50,000 referring expressions generated from this game. We are going to use this data set to train and evaluate our speaker and listener models. An example of what this data looks like is in Figure 1.

To obtain the non-contextual color representations the DNC is trained with, we are going to use another dataset of single colors and captions. These data have been collected by (Munroe, 2010) and cleaned by (McMahan and Stone, 2015). In these data collecting scenarios, human participants just had to label individual colors outside of a reference game scenario. There were approximately 200,000 participants labeling about 5 million colors, which provides a reasonable amount of data for training.

3.2 Models

- **Speaker-Listener Model (no common ground)**: Our first model is derived from the models of (Monroe et al., 2017). It involves two components: a literal listener and a literal speaker. The literal speaker acts as a conditional language model. It principally consists of two LSTMs. The first LSTM runs over the colors, with the target color last and produces a representation of all the colors. The second LSTM is the language modeling component—it is trained to predict the next token in the description based on previous tokens and the color encoding. The literal listener does the reverse: given an utterance, it runs its own LSTM over the utterance to produce a mean vector, μ, and co-variance matrix, Σ, that it uses to score each potential color, c:

  $$\textbf{Score} = (c - \mu)^T \Sigma (c - \mu)$$

  A softmax is then computed over the scores to select the color that is most likely target.

  Together, these models are assessed by playing the reference game: the literal speaker is presented with a context with the target color explicitly labeled, and creates an utterance. The literal listener then has to interpret this utterance by greedily sampling the most probable next token (essentially performing beam search with a beam size of 1). Their joint success at communicating the correct color forms the basis of their assessment. The Literal Speaker was trained for 30 epochs with the aid of GloVe embeddings (Pennington et al., 2014), and the Literal Listener was trained for 5 epochs. The 54-dimensional Fourier transform color representations of the colors, as discussed in (Monroe et al., 2017), were used.

  A visual representation of this architecture is visible in Figure 2.

- **DNC Model (with common ground)**: We hypothesize that by that by including a differentiable neural computer that stores color information as part of the inference procedure, we will be able to beat the performance of our Speaker-Listener model at this task. The DNC will be included as follows. First, using the dataset provided by (McMahan and Stone, 2015), the DNC will be trained with an autoencoder objective. We will use an encoding of the color as the read and write keys and an encoding of the color name will be stored in the DNC. This color name encoding will then be used to attempt to recreate the original color used as the key. These will be trained on the non-contextual color captions.

  At inference time, while playing the contextual color reference game, the DNC will be be queried with each of the colors presented and the representations extracted will be used as the color representations in the literal listener and speaker, rather than the raw color values (or their Fourier transforms as (Monroe et al., 2017) uses).

  In a sense, we are testing a portion of the proposition that Graves et al. (Graves et al., 2016) put forth in their Merlin system. They use a variational autoencoder and a DNC to learn how to store compact state representations that help their agent perform well on a number of tasks without explicit training. By training our DNC on an autoencoding objective, we hope that the the DNC will learn to store representations that will help color-game playing agents in their downstream contextual task. The DNC was trained for 5 epochs, once with 3-dimensional RGB color representations and once with 54-dimensional Fourier transform representations of the colors as detailed in (Monroe et al., 2017). Note that we had to train each model multiple times before obtaining our final models, and each training took almost an entire day, so we were not able to train as many models as we would have liked.

4 Results and Discussion

To score our models, we feed each the speaker model a color context (set of three colors) from the dataset, obtain a natural language utterance, which we feed as input to the listener. If the listener then selects the correct color, we consider this a correct answer. Our accuracy metric for the overall joint model is the rate at which the listener model selects the correct color.

	Joint Accuracy	Listener-Only Accuracy
Human Speaker and Listener	0.90*	—
Speaker-Listener	**0.7304**	**0.7438**
DNC Model, RGB Colors	0.3231	0.3333
DNC Model, Fourier Colors	0.3334	0.3335

Table 1: Joint accuracy (listener accuracy when paired speaker, which is given the color context) and listener-only accuracy (listener accuracy when given the color context and the corresponding human-produced utterance) for both the joint Speaker-Listener model and the DNC Model on the test set, as well as the human accuracy for reference. Bolded values are the best-performing values. *Human accuracy was only given in (Monroe et al., 2017) to two decimal places.

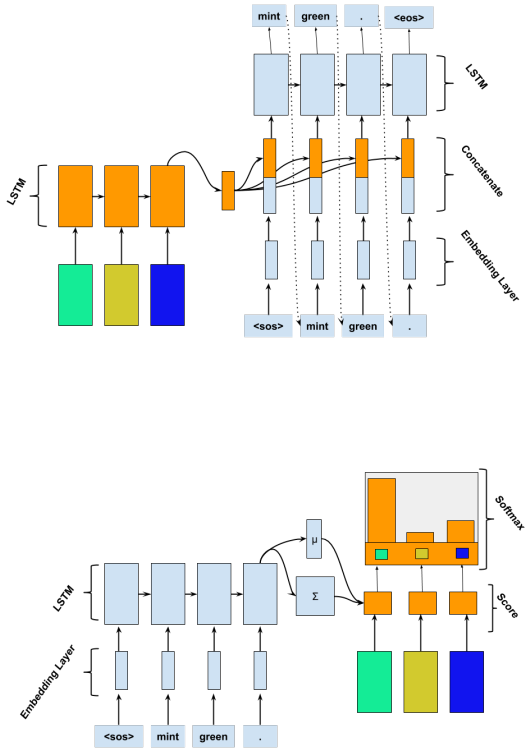

Figure 2: Speaker-Listener model architecture with Literal Speaker above and Literal Listener below.

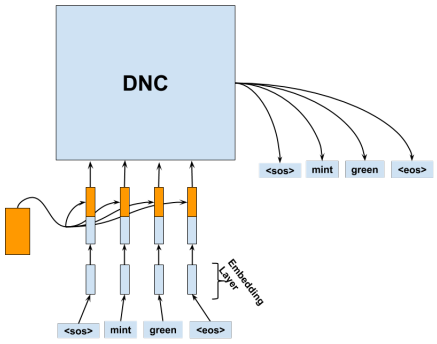

Figure 3: DNC model architecture. The DNC Speaker Model acts as a language model.

4.1 Overall Results

The performance of both our Speaker-Listener model and the DNC Model are shown in Table 1. We calculate the accuracy of the models where both our speaker and listener models are used, as well as the listener-only accuracy, or the accuracy achieved when only our listener model is used and its input is instead the human utterances directly from the dataset. Contrary to our hypothesis, the Speaker-Listener model performs much better than the models that incorporate a DNC.

First, the joint accuracy of our Speaker-Listener model using both the speaker and the listener models we implemented is only marginally less than the listener-only accuracy. This means that the speaker model we implemented performs well at producing informative utterances that are similar in quality to those produced by the humans in the dataset.

Turning to the DNC Model using RGB (3-dimensional) representations, the joint accuracy and listener-only accuracy were both approximately one-third. Though this is unexpected, it is explainable. Upon further investigation into the representations of colors that the DNC model constructed, we noticed that every sequence of tokens generated by the DNC, regardless of the color input, was the start token followed by the end token. We posit that the reason for this degeneracy is that there is insufficient signal in the RGB representations of the colors to learn complex mappings from tokens to colors. This means that the listener model effectively performed random guessing on the choices it was given, which comes out to around $\frac{1}{3}$ because there are three color choices per context.

Finally, for the DNC Model with Fourier color (54-dimensional) representations, we also achieve a similar accuracy level that amounts to random

guessing. While the tokens generated by the DNC queries were not simply start and end tokens, the mappings learned were rather nonsensical and did not contain intuitive words that described the colors. More importantly the representations produced for all colors were identical–consisting of the single token "electric". Since the listener was trained on human utterances, it makes sense that the DNC speaker did not offer any informative information about the colors, resulting in random guesses from the listener. This would be intuitively equivalent to having the listener select from three squares that are all the same color, which would result in random guessing.

We suspect that the poor performance of the DNC Model, both with RGB and with Fourier color representations, may be due to the amount of epochs for which the DNC was trained. Since the DNC was extremely expensive to train, we were only able to train it for 5 epochs. Even though the loss decreased significantly from the first epoch to the fifth (1.208 to 0.2493), and there should have been enough signal in the Fourier transform representations of the colors, the DNC likely just didn't see the colors frequently enough to create meaningful representations. For future work, we would like to explore training the model much more to see if it can avoid degeneracy and better model common ground.

4.2 Speaker Model Caption Generation without Common Ground

The Speaker-Listener model performed quite well in terms of accuracy without any common ground modeling, and we would like to highlight some examples of color contexts and corresponding captions generated by the speaker in Figure 4. In the first and second contexts, we have three colors that are very similar. However, the model still produces an utterance that exhibits some level of distinguishing between the colors.

In the first context, we see that the first color is more purple and the other two are more pink, and the model outputs *purple*. Similarly, in the second context, the first color has the largest hint of blue in it, and the model outputs *blue* to distinguish it from the others. It's especially interesting that the human speaker in this case also chose the exact same utterance.

In the third context, we have two colors that are relatively similar (the two greens)—impressively,

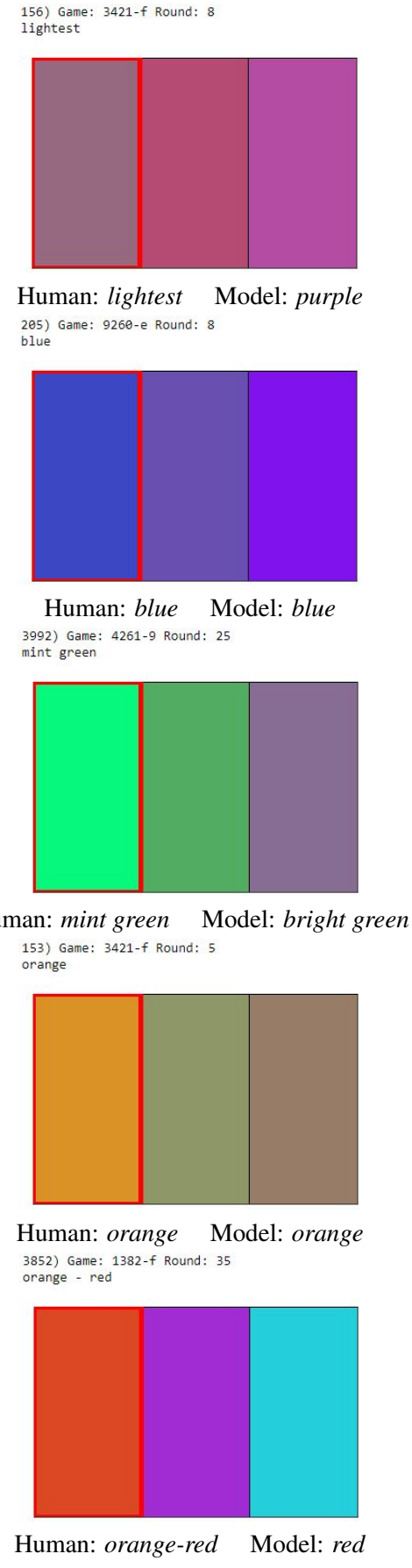

Figure 4: Examples of color contexts and captions generated by the speaker in the Speaker-Listener model, along with corresponding human utterances.

the model qualifies the green by outputting *bright green* to distinguish it from the duller green. The human speaker chooses to qualify the green with *mint* instead, but the intent is similar in that both recognize the need to distinguish the target from another similar distractor.

In the fourth and fifth contexts, the target color is clearly distinguishable from the other two colors. For the fourth context, the model behaves as expected and outputs the same utterance as the human speaker—*orange*. For the fifth context, the model outputs *red*, which is accurate and sufficient to distinguish between this color and the others. Interestingly, however, the human speaker actually gives more qualifying information than the model and says *orange-red* despite the lack of other oranges or reds in the context. This is likely because the human has ungrounded knowledge of the canonical "red" that English speakers are most familiar with, and since this color doesn't necessarily align perfectly with that prior knowledge, the human speaker felt the need to qualify the statement. The model, on the other hand, has learned to use the most efficient statement it needs to complete the task, and in this context, there is no need for qualification. It also does not have any form of common ground, which can also contribute to this bluntness.

In examining all of these examples, it is clear that the model has learned a mapping from utterances to colors and has learned some level of disambiguation between similar colors when they are present in the same context. Quite promising is the similarity between the utterances produced by the model and by the human speaker. Future work in modeling the impact of ungrounded knowledge on verbosity, of the sort illustrated in the fifth example in Figure 4, would also be interesting.

5 Conclusion and Future Work

In this work, we explored two ways to model agents in reference games using neural networks: a Speaker-Listener neural network with LSTM networks for both the speaker and listener, as well as a similar joint network that uses a differentiable neural computer to act as a language model that models common ground between the speaker and listener. We used the Colors in Context dataset (Monroe et al., 2017), from which we obtained color representations of the colors in the reference game. For the DNC, we used the non-contextual color descriptions dataset (Munroe, 2010) to model prior ungrounded knowledge of colors. While the DNC Model did not perform as well as the Speaker-Listener model, we were able to achieve pretty impressive accuracies with the Speaker-Listener model, which also had reasonable natural language utterances from the speaker that resulted in a listener accuracy similar to that achieved when human utterances were used.

In terms of future work, we see a lot of room for improvement upon this framework for modeling common ground. First, we do hope to be able to train the DNC for many more epochs to give the model a fair chance to learn proper representations of colors. We also hope to explore using other color representations and speaker and listener architectures to model common ground. Another exciting potential area of exploration would be to jointly train a speaker and listener model end-to-end, rather than separately as done in this work, to see if the common ground can be utilized even better. Constraints on the models would need to be applied so that the intermediate representations are intelligible natural language utterances and thus interpretable. This might also increase the accuracy of the models and bring the models closer to performing at the level of human accuracy. Finally, as mentioned, more work on modeling the impact of ungrounded knowledge—both helpful and potentially over-informative—on speaker utterances in contextual tasks is also of interest.

Acknowledgements

As we used similar datasets in another collaborative effort that also used the literal speaker implementation, we would like to thank our collaborator for that project, Suvir Mirchandani, for his assistance in developing the literal speaker sampling methodology.

References

Jacob Andreas and Dan Klein. 2016. Reasoning about Pragmatics with Neural Listeners and Speakers. Technical report.

Reuben Cohn-Gordon, Noah Goodman, and Chris Potts. Pragmatically Informative Image Captioning with Character-Level Inference. Technical report.

Vladan Devedzic. 2001. Knowledge modeling–state of the art. *Integrated Computer-Aided Engineering*, 8(3):257–281.

Noah D Goodman and Michael C Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829.

Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. 2016. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471.

Justine T Kao, Jean Y Wu, Leon Bergen, and Noah D Goodman. 2014. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, 111(33):12002–12007.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. *arXiv preprint arXiv:1805.04623*.

Fereshte Khani, Noah D Goodman, and Percy Liang. 2018. Planning, inference and pragmatics in sequential language games. *Transactions of the Association for Computational Linguistics*, 6:543–555.

Mike Lewis, Denis Yarats, Yann N Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125*.

Brian McMahan and Matthew Stone. 2015. A bayesian model of grounded color semantics. *Transactions of the Association for Computational Linguistics*, 3:103–115.

Will Monroe, Robert XD Hawkins, Noah D Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Randall Munroe. 2010. Color survey results.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware Captions from Context-agnostic Supervision. Technical report.