Developing a Pragmatic Framework for Evaluating Color Captioning Systems

Benjamin Newman **Suvir Mirchandani** **Julia Gong**
{ blnewman, smirchan, jxgong } @stanford.edu

Abstract

In this paper, we present a framework for evaluating natural language descriptions in the color captioning problem. In this task, two agents are given a set of three colors and one of them generates a description of a target color for the other agent. Our approach is pragmatically motivated: we measure the effectiveness of a caption in terms of how well a trained model can select the correct color given the caption. We investigate four models, two of which explicitly model pragmatic reasoning, and we formulate a performance metric based on Gricean maxims to compare the effectiveness of the models. Our results indicate that though modeling pragmatic reasoning explicitly does improve evaluation performance by a small margin, it may not be essential from a practical perspective. Overall, we believe this evaluation framework is a promising start for evaluating natural language descriptions of captioning systems.

1 Introduction

Recently, large natural language generation models like OpenAI's GPT-2 have shown an impressive ability to produce human-like utterances. Despite these successes, systematically and thoroughly measuring the performance of these models remains elusive.

Currently, there are three main approaches to evaluate model generated utterances. The most direct option is to ask humans to judge the quality of generated content (Hashimoto et al., 2019). While this is quite effective for assessing language quality, the cost is prohibitive at larger scales. A second approach is to use n-gram overlap metrics that count the number of n-grams that appear in both a generated utterance and a reference one. There are a variety of these scores, such as BLEU, METOER, and CIDEr, and each is used in particular sub-fields (e.g. translation, summarization, and image captioning) (Vedantam et al., 2015). These metrics are somewhat unsatisfying, however, because they depend on the tokens rather than some deeper semantic notion. More significantly, these scores do not correlate particularly well with human judgments (Novikova et al., 2017). A third option is to try to *learn* a metric that produces a score for utterances that correlates with human judgments (Cui et al., 2018). These metrics can be difficult to train reliably and still require human annotations indicating the quality of machine generated utterances.

In the following work, we propose a perspective on this evaluation problem based in grounded language understanding. Our approach is rooted in the pragmatic notion that a good utterance is one that allows another agent to achieve a specific task. Here, we focus on the Colors in Context task proposed by Monroe et al. (2017). In this reference game, two agents—a *listener* and a *speaker*—cooperate to distinguish a target color from two other distractor colors: the speaker produces an utterance to identify the target and the listener selects the color they believe is the target based on the speaker's description (example in Figure 2). Taking inspiration from the third approach to metrics—creating a model to evaluate utterances—we implement a number of models that try to distinguish between good and bad speaker captions for describing a target color to a listener. The main difference in our approach is that we claim the quality of the caption is based on the ability of our models to perform the task rather than predict human ratings of quality.

The core motivation behind our work is that a good caption is one that allows an average listener to identify the color that the speaker is describing. In order to assess computational models of speakers, we first seek a model that assigns higher scores to better speakers and lower scores

to worse speakers. Because Monroe et al. (2017) have shown that explicitly pragmatic agents perform better at this task than non-pragmatic agents, we hypothesize that a metric that explicitly models pragmatics will correlate better with human performance than a metric that does not explicitly take pragmatics into account. Our pragmatic modeling takes two forms. The first is imagination—the evaluating model tries to recreate the target that the speaker sees given the context and the caption. The second is the recursive reasoning of Goodman and Frank (2016) that has shown a lot of promise for modeling pragmatic scenarios.

2 Related Work

Much of the work related to ours comes from the fields of grounded language understanding and evaluation metrics for generated text, the latter of which was discussed previously.

2.1 Grounded Language Understanding

The field of grounded language understanding mainly concerns itself with using language to interact with the world. The grounded reference games that are ubiquitous in the literature can be used to model a host of linguistic behavior, from negotiation dialogues to hyperbole (Lewis et al., 2017; Kao et al., 2014). The Rational Speech Acts (RSA) framework proposed by Goodman and Frank (2016) explicitly encodes the kind of recursive reasoning that improves reference game performance. Essentially, players reason about one another's hypothetical actions when producing utterances and making choices. Most of these games involve hand-crafted lexicons, but Andreas and Klein (2016); Vedantam et al. (2017) and Mao et al. (2016) extend these notions to more open-domain vocabularies and realistic images and scenarios.

2.2 Color Understanding

A previous attempt at producing captions for colors was undertaken by McMahan and Stone (2015). They create a Bayesian model called LUX to caption single colors in a non-contextual and non-reference game scenario. Monroe et al. (2016) improve upon their captions by bringing in long short-term memory recurrent neural networks (LSTMs). The data for a color reference game was also collected and analyzed by Monroe et al. (2017), as the way colors are described often varies based on the colors around them. In their work, a number of methods similar to those of Andreas and Klein (2016) are used to create agents to model speakers and listeners. With all of this work done on designing and building these agents, there has not to our knowledge been work that specifically uses these techniques for evaluation.

3 Methods and Metrics

We implement four caption evaluation models, as well as a performance and correlation metric, to compare their effectiveness. We refer to the evaluation models as *listener models*: given a caption and color context (the set of three colors), they assign a score to each of the colors, just as a listener in might implicitly do when selecting a color in a reference game. We use the score that a listener model assigns to the target color in a color context as that listener model's score for the accompanying caption.

3.1 Feature Representation

Colors are represented as three-dimensional vectors in RGB space, which are then transformed into a Fourier basis representation, as done by Monroe et al. (2016). (The Literal Speaker instead uses HSV as the initial color representation because we find it performs better with HSV). The Fourier representation captures periodicity in the base color spaces. We represent captions using token indices after applying the same preprocessing procedure as Monroe et al. (2017).

3.2 Models

Three of the four models (the Baseline Listener, the Literal Listener, and the Pragmatic Listener) are probabilistic, while the Imaginative Listener outputs a color distance metric. The Literal Listener and Pragmatic Listener are based heavily on (Monroe et al., 2017). The Pragmatic Listener involves a submodel—the Literal Speaker—which we discuss further below. We illustrate the four model architectures in Figure 1.

3.2.1 Baseline Listener

We first implement a two-layer baseline network (Figure 1a), which ignores the caption entirely and only uses the color inputs to predict the target color. We expect this model to be effectively equivalent to educated random guessing. It is still useful to observe this to confirm that the other evaluation models are picking up on signal from

the caption in tandem with the color context, rather than simply predicting the target color from the color prompt alone.

To use the Baseline Listener as an evaluation model for captions, we output the probability corresponding to the target color.

3.2.2 Literal Listener

Our Literal Listener model is derived from the Base Listener agent of (Monroe et al., 2017). It runs a bidirectional LSTM over the utterance to predict a Gaussian distribution over colors, parameterized by a mean vector, μ and covariance matrix Σ. The caption tokens are embedded in a 100-dimensional input space, and the LSTM has 100 hidden dimensions. The output distribution is sampled at each color representation c in the context to produce a score of the form

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

The scores are normalized using a softmax function to produce a probability distribution, as shown in Figure 1b. Like the Baseline Listener, we extract the probability assigned to the target color and use this as the Literal Listener's model score.

3.2.3 Imaginative Listener

Next, we develop an Imaginative Listener (Figure 1c), which attempts to directly predict the target color given the caption and two distractor colors. Captions are embedded in a 100-dimensional input space (initialized with GloVe embeddings) and are passed through a bidirectional LSTM. The color representations are embedded using a linear layer, and the two intermediate representations are concatenated and passed through two linear layers.

The output is a color in RGB space that represents the color the listener model predicts is the target indicated by the speaker. To convert this to a usable model score, we use the CIEDE 2000 perceptual color difference (Luo et al., 2001) between the predicted color and target color.

3.2.4 Pragmatic Listener

Finally, we implement a Pragmatic Listener (Figure 1d), which models recursive reasoning between a hypothetical speaker and listener, based on Monroe et al. (2017) and the Rational Speech Acts model (Goodman and Frank, 2016).

The Pragmatic Listener reasons about the possible behaviors of a Pragmatic Speaker, which in turn reasons about the possible behaviors of the Literal Listener. More precisely, let the Literal Listener be modeled as $L_0(t|u, C; \theta)$ where t is a color, u is an utterance, C is a color context, and θ are the learned weights (L_0 is learned by the model described in Section 3.2.2).

The Pragmatic Speaker distribution $S_1(u|t, C; \theta)$ is determined by L_0:

$$S_1(u|t, C; \theta) = \frac{L_0(t|u, C; \theta)^\alpha}{\sum_{u' \in U} L_0(t|u, C; \theta)^\alpha}$$

where α is a parameter controlling the degree of pragmaticism of the speaker. Loosely, the Pragmatic Speaker weights an utterance based on how likely the Literal Listener would respond correctly to that utterance and normalizes across all utterances in the universal set U.

The Pragmatic Listener model performs similar reasoning about the Pragmatic Speaker:

$$L_2(t|u, C; \theta) = \frac{S_1(u|t, C; \theta)}{\sum_{t' \in C} S_1(u|t', C; \theta)}.$$

In order to approximate the normalization in the Pragmatic Speaker (which cannot be performed directly, because it sums $L_0(t|u', C; \theta)$ over all $u' \in U$), we implement a submodel—the Literal Speaker—to provide plausible captions over which the Literal Listener can be sampled. This is the same approach that Monroe et al. (2017) take.

Literal Speaker The Literal Speaker encodes a color prompt and generates a caption for the target color. It consists of two LSTMs. One LSTM runs over the colors, with the target color last and produces a representation of 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.

To construct a limited universe U' of plausible captions for the Literal Listener, we randomly shuffle the color inputs and then iterate through the three colors as hypothetical targets for the Literal Speaker to describe. For each color, we generate k sample captions and add it to U'. The k samples are chosen using beam search (where depth corresponds to the number of tokens) to approximate the top-k predictions of the Literal Speaker. The denominator in the Pragmatic Speaker equation then sums over all $u \in U'$ rather than the set of all utterances U.

The Pragmatic Listener performs Bayesian inference on the Literal Listener outputs as described above to determine a context-relevant

color prediction. For this model to serve as a caption evaluation model, we score it using the probability it assigns to the true target color.

3.3 Performance Metrics

To evaluate our models' performance, we introduce the True Score and Gricean True Score (GTS), speaker performance metrics based on the rate of correct target identification by listeners in the synthetic dataset. The score S_{Grice} is formulated as $S_{Grice} = \frac{\overline{s}}{\overline{t} * \overline{n}}$, Where $\overline{s}$ is the True Score, or the average rates of correct target identification in each 50-round game in the dataset, $\overline{t}$ is the list of average listener click times in each game, or the time stamp of the listener's clicking on the color to end the round, and $\overline{n}$ is the average number of words in the speaker's utterance for each game.

The intuition behind the Gricean True Score is that it augments the simple True Score by encoding the importance of speaker efficiency and accuracy. We expect speakers with higher scores to exhibit a positive relationship with correct target identification by listeners, so the average accuracy is proportional to S_{Grice}. Conversely, higher scores should be inversely proportional to the time it takes the listener to process their utterance (assuming a competent listener), as well as the length (inefficiency) of the speaker's utterance.

To evaluate our models, we then formulate the GTS correlation metric using the Gricean True Score. This metric is the Pearson correlation between the per-game average target identification accuracies of a given model and the Gricean True Scores for the aggregate data (not split by condition). However, as will later be presented, it is still informative to also investigate the individual performance metrics of each model on the *close*, *split*, and *far* conditions separately.

4 Data

4.1 Color Reference Dataset

To develop our evaluation framework, we use the Color Reference dataset (Monroe et al., 2017).[1] Their corpus was created using 967 participants on Mechanical Turk who played a total of 1,059 reference games with 50 rounds each.

In each game, a participant was assigned to be either a speaker or a listener. The task of the speaker was to communicate which of three colors was a specified target color; the listener would guess the target from the same set of three colors.

The trials were split evenly among three conditions for the color contexts: *far*, *split*, and *close*. In the far condition, the three colors were far apart in RGB color space; in the split condition, the target color was nearby to exactly one of the other colors; and in the close condition, all three colors were close in color space. Figure 2 shows three context-caption pairs, one for each condition.

After data cleansing, the dataset consists of 948 games across 46,994 rounds and 53,365 speaker utterances. The split provided by Monroe et al. (2017) has 15,665, 15,670, and 15,659 entries in the training, development, and testing sets, respectively. For each entry sent, the dataset provides, among other fields, the following information: game identifier, round number, worker identifier, round condition, time of message, the three colors and their positions, which color was the speaker's target, which color was selected by the listener, and time of listener's click.

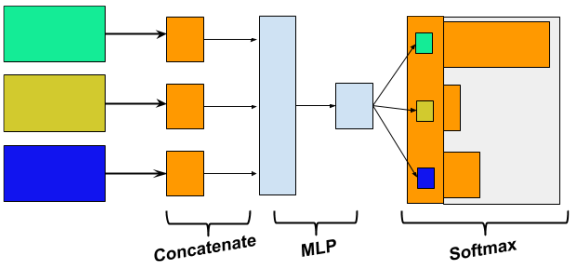
To train our listener and speaker models, we ignore listener messages and concatenate speaker messages within each round.

4.2 Synthetic Data

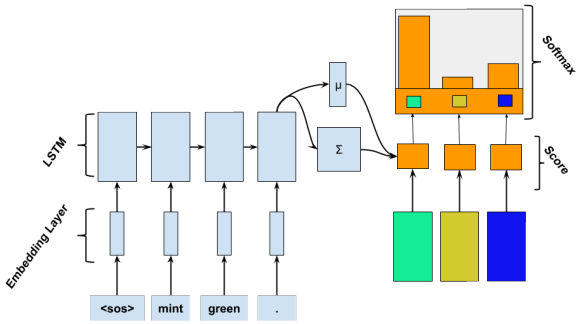
As explained, our goal is to develop listener models that can evaluate speaker model captions. To do this, we want to compare the human listeners' choices based on captions in the dataset to our listener models' choices based on these same utterances. Our evaluation method thus relies on the strength of the correlation between these values. The issue with the Color Reference dataset as given, however, is that 90% of games result in the human listener successfully identifying the target color (97% for far condition, 90% for split, and and 83% for close). This accuracy imbalance provides little signal to evaluate between models—since humans are very good at this task, the utterances they create are all of about equal quality even when the listener chooses the incorrect color.

To avoid both this issue, we use the raw dataset to construct a synthetic dataset that includes poor utterances. To create bad speaker utterances, we alter the utterances in rounds that the listener got correct by changing the target color the speaker is referring to but keeping the caption. For the far and close conditions, we consider both possible

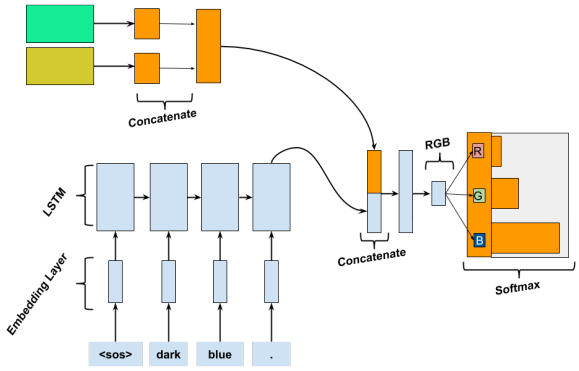
[1]The dataset is open-source and is available at https://cocolab.stanford.edu/datasets/colors.html.



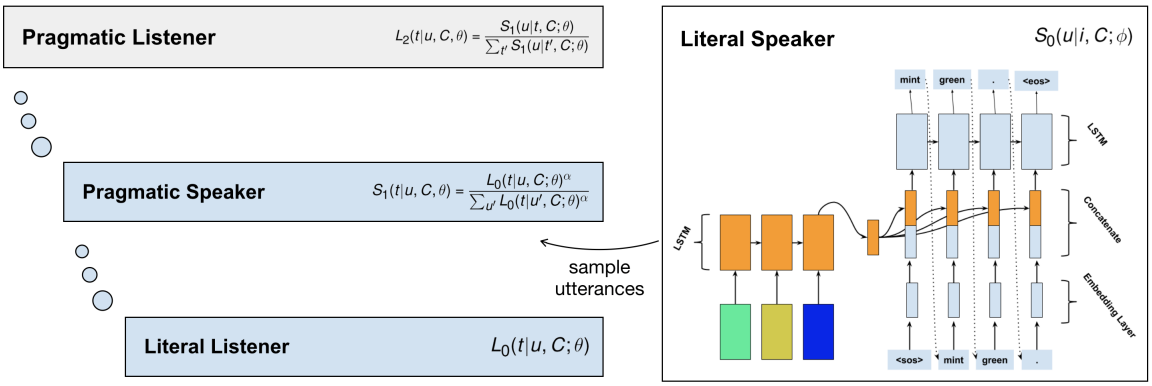
(a) Baseline Listener



(b) Literal Listener



(c) Imaginative Listener



(d) Pragmatic Listener

Figure 1: Model architectures for the (a) Baseline Listener, (b) Literal Listener, (c) Imaginative Listener, and (d) Pragmatic Listener.

Figure 2: Example colors and captions for each of the three color conditions in the Color Reference dataset.

distractors as targets, while in the split condition we only consider the closest distractor. From the speaker's perspective, the utterance they produce is uncooperative and misleading—therefore, a bad utterance. From the listener's perspective, nothing has changed, and the listener takes the same action, selecting the same color they did before the target was switched. This is now an incorrect color choice.

Out of a mix synthetic and unaltered utterances we construct "speakers" of varying quality. Each speaker makes 50 utterances (the length of one of the reference games). There are an approximately equal number utterances in each of the close, far and split conditions. 11 types of synthetic speakers are created, each with a certain percentage of correct utterance ranging from 0 to 100%. The correct and incorrect utterances are also spread as evenly as possible across the three conditions. There are 47 of each type of speaker for a total of 517 synthetic speakers.

It is important to note that this synthetic dataset is only used for *evaluation*, so we create one from the development set and one from the test set. This data is not used for training—the developed metrics are based on the principles of pragmatics, which depend on the speaker being cooperative. The speakers in the synthetic data are certainly not cooperative. This behavior makes them unpredictable and therefore more difficult for the metric models to learn.

5 Results and Discussion

5.1 Model Hyperparameter Tuning

In order to determine the best hyperparameter settings for each of our models, we perform a grid search for each listener model, using the average probability the model places on the target color across the validation set to score the model's performance. We do not do this for the baseline model because it is already performing at the level of guessing we expect it to be able to achieve.

For the Literal Listener, we try learning rates $lr \in \{0.0005, 0.001, 0.004, 0.01\}$ and LSTM hidden dimension sizes $h_{lstm} \in \{50, 100, 150, 200\}$. For the Imaginative Listener, we try the same set of learning rates and hidden dimension sizes, as well as the color embedding hidden dimension size $h_{col} \in \{50, 100, 150, 200\}$ and w, whether or not we use GloVe vectors in the weight matrix. For the Pragmatic Listener, we try the values $\{0.5, 1, 2\}$ for the α parameter. For the Literal Speaker submodel, we use the hyperparameters identified by Monroe et al. (2017).

We find that the following combinations of hyperparameters result in the highest score on the development set for our listener models:

- Literal Listener: $lr = 0.0005$, $h_{lstm} = 100$
- Imaginative Listener: $lr = 0.001$, $h_{lstm} = 50$, $h_{col} = 50$, $w = \texttt{True}$
- Pragmatic Listener: $\alpha = 0.5$

5.2 Performance Metrics

To calculate the GTS correlation metric for each of our models, we choose the best hyperparameter setting for each of our models and train each model 10 times, each time with a random initialization, and run each of these models on the test set and calculate their correlation metric. (For the Pragmatic Listener, we keep the the trained submodels fixed to express the randomness of the Pragmatic Listener through the sampling procedure.) Here, we present the average GTS correlation metrics of these trials for each model for each of the close, split, and far conditions in Table 1 and Figure 2, as well as in aggregate (the aggregate data not split by condition). We also include the corresponding margins of error given these repeated trials.

In addition, we report the accuracy of the Imaginative Listener in terms of mean color distance rather than whether the closest color is the target. This is because in the split and close conditions, all of the colors are close, so it is more difficult for the closest to be the target, whereas in the far condition, the generated color just has be distant enough from the distractors to be closest to the target.

To provide a more visual examination of each model's score correlation patterns, we also plot the model scores against the Gricean True Scores and True Scores for each condition in Figure 3.

Across the board, as expected, the close and split conditions have lower correlations than the far condition. In addition, the Pragmatic Listener edges out the Literal Listener by a small, yet significant margin, which performs better than the Imaginative Listener. The Imaginative Listener's worse performance makes sense because it needs to synthesize, as opposed to select, the color. The latter result also supports our hypothesis that

	$\rho_{aggregate}$	ρ_{close}	ρ_{split}	ρ_{far}	Accuracy / Distance
Baseline	0.0151 ± 0.0234	-0.0010 ± 0.0258	0.0022 ± 0.0307	0.0210 ± 0.0276	0.0201 ± 0.0177
Literal	0.9572 ± 0.0021	0.8647 ± 0.0078	0.8899 ± 0.0074	0.9467 ± 0.0014	$\mathbf{0.4599} \pm 0.0012$
Imaginative	0.8882 ± 0.0024	0.4474 ± 0.0087	0.5298 ± 0.0138	0.8664 ± 0.0087	24.0609 ± 0.0662
Pragmatic	$\mathbf{0.9617} \pm 0.0001$	$\mathbf{0.8845} \pm 0.0004$	$\mathbf{0.9047} \pm 0.0003$	$\mathbf{0.9486} \pm 0.0001$	0.4531 ± 0.0004

Table 1: Mean performance metrics (correlation metrics ρ) for the Baseline, Literal, Imaginative, and Pragmatic Listener models with 95% confidence intervals. Imaginative Listener correlations are negated for consistency, as discussed in Figure 4. Accuracy refers to mean target identification accuracy for all models except Imaginative Listener, where it refers to mean perceptual distance to target in color space. Because half of the synthetic test data is intentionally misleading, perfect model accuracy should be 50%. Bolded values are the best-performing values.

incorporating pragmatic reasoning into the Pragmatic Listener model boosts performance.

6 Discussion and Analysis

Overall, our models performed quite well and mostly aligned with our hypotheses.

The Baseline Listener performed at random. Note that random in this scenario is not 33%: in the split condition, the only two viable choices are the ones that are close to each other, so the expected performance is $\frac{1}{3}(50\%) + \frac{2}{3}(33\%) = 38\%$. This means our baseline was able to learn some notion of color similarity and then randomly guessed among equally plausible choices.

Next, the Imaginative Listener performed worst out of the three models we experimented with. This is contrary to our hypothesis that the Imaginative Listener would perform better than the Literal Listener because its architecture is pragmatically motivated. The relatively worse performance can be explained by the fact that Imaginative Listener does not have access to the target color and has to synthesize it, whereas the other models merely had to select it. Intuitively, the reason for the worse performance is similar to why a "fill-in-the-blank" question is more difficult than a multiple choice exam. Looking at the different conditions can provide further insight. We see that most of its correlation can be attributed to the colors generated in the far condition, which were very distant from the targets when the captions were poor. We can also see that regardless of the Gricean True Score of the speaker, the split and close conditions tended to lead to color predictions that were visually closer to the targets compared to the far condition. This is likely because the colors the model was conditioning on were more similar to the targets.

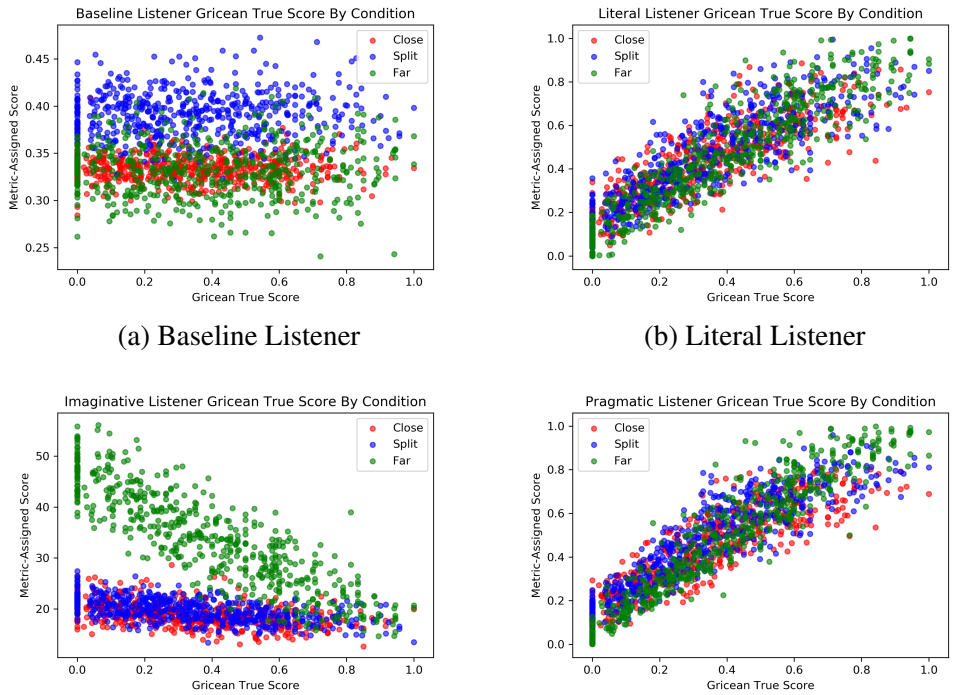
The lack of correlation between Gricean True Scores and Imaginative Listener model scores is likely due to the prevalence of negation and comparatives in the split and close conditions that do not appear in the far condition. Monroe et al. (2017) notes that in the more difficult conditions, human speakers rely more heavily on comparatives and negations to distinguish colors. This means there might be more distracting signal in the caption that the Imaginative Listener picks up on and converts to color. It could also mean that speakers rely more on adjectives for disambiguation without actually mentioning the colors—e.g., "the bright one" does not give the network much signal for the actual color that should be produced.

Finally, it's interesting to note that the Literal Listener and Pragmatic Listener performed similarly well, especially on the far condition. However, true to our hypothesis, of the differences we do see, the close and split conditions see the largest gains from the Pragmatic Listener, which is where we would expect the recursive reasoning to help. The correlations between the Gricean True Scores and the model scores were high for both, so both can serve as viable caption evaluation models. The similarity implies that the pragmatic sampling procedure did not greatly impact the performance of the Literal Listener. This is puzzling from a theoretical perspective, but good from an implementation perspective because the Pragmatic Listener is more computationally expensive.

The tight confidence intervals and similarity of the scores across random initializations is promising as well, as it implies that our models are likely finding equally good optima during training.

7 Conclusion and Future Work

This work explores a novel framework for evaluation of natural language descriptions in the color captioning problem. We present four models: a



(a) Baseline Listener

(b) Literal Listener

(c) Imaginative Listener

(d) Pragmatic Listener

Figure 3: Examples of model correlation patterns between model scores and Gricean True Scores for the various listener models. Note that the Imaginative Listener is expected to have negative correlation because the model score is the distance between the target and generated color; therefore, for evaluation, we negate this score.

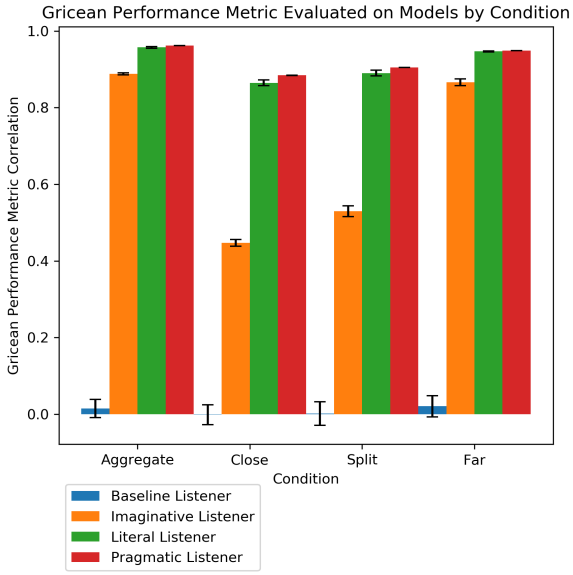


Figure 4: Model score correlations with Gricean True Scores across the different models and conditions. Each bar represents the average of ten trials and has error bars with 95% confidence intervals.

Baseline Listener, Literal Listener, Imaginative Listener, and Pragmatic Listener. We also present the Gricean True Score (GTS) correlation metric for evaluating model efficacy. We find that the Imaginative Listener does not perform as well as the Literal and Pragmatic Listeners. Moreover, as hypothesized, we find that modeling pragmatic reasoning in the Pragmatic Listener does offer improvement over the Literal Listener in terms of the GTS Correlation, though it is more computationally expensive due to its recursive inference. Overall, we believe the framework of using a performance metric like the GTS to evaluate captioning models is a promising direction for future research.

We envision multiple ways to extend this work. In terms of color captioning, there is room to explore better baseline models, performance metrics, and more advanced pragmatic reasoning models. In the bigger picture, we are interested in extending this framework to non-contextual settings and other grounded language tasks, such as image caption evaluation. This might take the form of using the caption to recreate salient features of an image, and a good caption would be one where these features can faithfully be recreated.

8 Acknowledgements and Contributions

We would like to thank Professor Christopher Potts and Reuben Cohn-Gordon of Stanford Linguistics for discussing ideas with us about the project direction.

B.N. set up the modeling framework, cleaned and processed data, and implemented the Literal Listener, Literal Speaker, and Imaginative Listener. S.M. implemented, tuned, and evaluated the Pragmatic Listener (including beam search for the Literal Speaker). J.G. implemented the Gricean True Score performance metric, and tuned and evaluated the Literal Listener. We wrote the paper and made the video collaboratively.

Appendix

The GitHub repository for our project code is here.

References

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Austin, Texas. Association for Computational Linguistics.

Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. 2018. Learning to evaluate image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5804–5812.

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

Tatsunori B Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. *arXiv preprint*.

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.

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*.

M Ronnier Luo, Guihua Cui, and B Rigg. 2001. The development of the cie 2000 colour-difference formula: Ciede2000. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 26(5):340–350.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

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, Noah Goodman, and Christopher Potts. 2016. Learning to generate compositional color descriptions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2243–2248, Austin, Texas. Association for Computational Linguistics.

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.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. *CoRR*, abs/1701.02870.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.