Reasoning about pragmatics with neural listeners and speakers

Jacob Andreas and Dan Klein
The reference game
The reference game
The reference game

The one with the snake
The reference game

Mike is holding a baseball bat
The reference game

bat a is holding Mike baseball
The reference game

They are sitting by a picnic table
The reference game

There is a bat
The reference game

There is a bat
The reference game

Why do we care about this game?

Don’t you think it’s a little cold in here?

Do you know what time it is?

Some of the children played in the park.
Mike is holding a baseball bat
Jenny is running from the snake
Mike is holding a baseball bat.
How to win

DERIVED STRATEGY: Reason about listener beliefs

DIRECT STRATEGY: Imitate successful human play

There is a snake

There is a snake

There is a snake
How to win

**DERIVED STRATEGY:**
Reason about listener beliefs

[Monroe and Potts, 2015]
[Smith et al. 2013]
[Vogel et al. 2013]
[Golland et al. 2010]

**DIRECT STRATEGY:**
Imitate successful human play

[Mao et al. 2015]
[Kazemzadeh et al. 2014]
[Fitzgerald et al., 2013]
How to win

DERIVED STRATEGY:
Reason about listener beliefs

**PRO:** pragmatics “for free”

**CON:** past work needs hand-engineering

DIRECT STRATEGY:
Imitate successful human play

**PRO:** domain repr “for free”

**CON:** past work needs targeted data
How to win

**DERIVED STRATEGY:**
Reason about listener beliefs

**DIRECT STRATEGY:**
Imitate successful human play

**Learn** base models for interpretation & generation without pragmatic context

Explicitly **reason** about base models to get novel behavior
Data

Abstract Scenes Dataset

1000 scenes
10k sentences
Feature representations
Approach

Literal speaker ➔ Sampler

Literal listener ➔ Reasoning speaker
A literal speaker \((S\emptyset)\)

Mike is holding a baseball bat
A literal speaker (**S₀**)
Module architectures

Referent encoder

Referent decoder
Training S0

Mike is holding a baseball bat
A literal speaker ($S\theta$)

Mike is holding a baseball bat
The sun is in the sky
Jenny is standing next to Mike
A literal listener (L0)

Mike is holding a baseball bat
A literal listener (L0)

Mike is holding a baseball bat

Descr. encoder

Referent encoder

Referent encoder

Scorer

0.87

0.13
Module architectures

Referent encoder

Referent decoder

FC
ReLU
Sum
Softmax
choice
Mike is holding a baseball bat

(random distractor)
A literal listener (L₀)

Mike is holding a baseball bat
A reasoning speaker (S1)

Mike is holding a baseball bat
A reasoning speaker (S1)

Literal speaker

Mike is a baseball bat

The sun is in the sky

Jenny is standing next to Mike

Literal listener
A reasoning speaker (S1)

Literal speaker

- The sun is in the sky: 0.05
- Mike is a baseball bat: 0.09
- Jenny is standing next to Mike: 0.08

Literal listener

- Mike is a baseball bat: 0.9
- The sun is in the sky: 0.5
- Jenny is standing next to Mike: 0.7
A reasoning speaker ($S1$)

- **Literal speaker**
  - Mike is a baseball bat
  - The sun is in the sky
  - Jenny is standing next to Mike

- **Literal listener**
  - $0.9^{1-\lambda} \times 0.05^\lambda$
  - $0.5^{1-\lambda} \times 0.09^\lambda$
  - $0.7^{1-\lambda} \times 0.09^\lambda$
Experiments

The hot air balloon is in the sky.

Which image does this caption describe?

- Left
- Right

Submit
Baselines

- **Literal**: the $L_0$ model by itself
- **Contrastive**: a conditional LM trained on both the target image and a random distractor

[Mao et al. 2015]
Results (test)

- Literal: 64%
- Contrastive: 69%
- Reasoning: 81%
Figure 5: Tradeoff between speaker and listener models, controlled by the parameter $\lambda$ in Equation 8. With $\lambda = 0$, all weight is placed on the literal listener, and the model produces highly discriminative but somewhat disfluent captions. With $\lambda = 1$, all weight is placed on the literal speaker, and the model produces fluent but generic captions.

4.1 How good are the base models?

To measure the performance of the base models, we draw 10 samples $d_{jk}$ for a subset of 100 pairs $(r_1, j, r_2, j)$ in the Dev-All set. We collect human fluency and accuracy judgments for each of the 1000 total samples. This allows us to conduct a post-hoc search over values of $\lambda$: for a range of $\lambda$, we compute the average accuracy and fluency of the highest scoring sample. By varying $\lambda$, we can view the tradeoff between accuracy and fluency that results from interpolating between the listener and speaker model—setting $\lambda = 0$ gives samples from $p_L$, and $\lambda = 1$ gives samples from $p_S$.

Figure 5 shows the resulting accuracy and fluency for various values of $\lambda$. It can be seen that relying entirely on the listener gives the highest accuracy but degraded fluency. However, by adding only a very small weight to the speaker model, it is possible to achieve near-perfect fluency without a substantial decrease in accuracy. Example sentences for an individual reference game are shown in Figure 5; increasing $\lambda$ causes captions to become more generic.

For the remaining experiments in this paper, we take $\lambda = 0.02$, finding that this gives excellent performance on both metrics.

On the development set, $\lambda = 0.02$ results in an average fluency of 4.8 (compared to 4.8 for the literal speaker $\lambda = 1$). This high fluency can be confirmed by inspection of model samples (Figure 4). We thus focus on accuracy for the remainder of the evaluation.

4.2 How many samples are needed?

Next we turn to the computational efficiency of the reasoning model. As in all sampling-based inference, the number of samples that must be drawn from the proposal is of critical interest—if too many samples are needed, the model will be too slow to use in practice. Having fixed $\lambda$ in the preceding section, we measure accuracy for versions of the reasoning model that draw 1, 10, 100, and 1000 samples. Results are shown in Table 1. We find that gains continue up to 100 samples.

4.3 Is reasoning necessary?

Because they do not require complicated inference procedures, direct approaches to pragmatics typically enjoy better computational efficiency than derived ones. Having built an accurate derived speaker, can we bootstrap a more efficient direct speaker? To explore this, we constructed a "compiled" speaker model as follows: Given reference candidates $r_1$ and $r_2$ and target $t$, this model produces embeddings $e_1$ and $e_2$, concatenates them together into a "contrast embedding" $[e_t, e_t]$, and then feeds this whole embedding into a string decoder module. Like $S_0$, this model generates captions without the need for discriminative rescoring; unlike $S_0$, the contrast embedding means this model can in principle learn to produce pragmatic captions, if given access to pragmatic training data. Since no such training data exists, we train the compiled model on (a) target (b) distractor

(a) target
(b) distractor
(prefer $L_0$)
0.0
a hamburger on the ground
0.1
mike is holding the burger
(prefer $S_0$)
0.2
the airplane is in the sky

Figure 5: Captions for the same pair with varying $\lambda$. Changing $\lambda$ alters both the naturalness and specificity of the output.
How many samples?

Accuracy

# Samples
Examples

the sun is in the sky
Examples

the dog is standing beside jenny
Examples

- The sun is in the sky
- The plane is flying in the sky
- The dog is standing beside Jenny
- Mike is wearing a chef’s hat

Figure 4: Four randomly-chosen samples from our model. For each, the target image is shown on the left, the distractor image is shown on the right, and description generated by the model is shown below. All descriptions are fluent, and generally succeed in uniquely identifying the target scene, even when they do not perfectly describe it (e.g. (c)). These samples are broadly representative of the model’s performance (Table 2).

Table 2: Success rates at RG on abstract scenes. “Literal” is a captioning baseline corresponding to the base speaker $S_0$. “Contrastive” is a reimplementation of the approach of Mao et al. (2015). “Reasoning” is the model from this paper. All differences between our model and baselines are significant ($p<0.05$, Binomial).

<table>
<thead>
<tr>
<th># of differences</th>
<th>Literal ($S_0$)</th>
<th>Reasoning</th>
<th>Compiled ($S_1$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50%</td>
<td>64%</td>
<td>44%</td>
</tr>
<tr>
<td>2</td>
<td>66%</td>
<td>86%</td>
<td>72%</td>
</tr>
<tr>
<td>3</td>
<td>70%</td>
<td>88%</td>
<td>80%</td>
</tr>
<tr>
<td>4</td>
<td>78%</td>
<td>94%</td>
<td>80%</td>
</tr>
<tr>
<td>Mean</td>
<td>66%</td>
<td>83%</td>
<td>69%</td>
</tr>
</tbody>
</table>

Table 3: Comparison of the “compiled” pragmatic speaker model with literal and explicitly reasoning speakers. The models are evaluated on subsets of the development set, arranged by difficulty: column headings indicate the number of differences between the target and distractor scenes.

<table>
<thead>
<tr>
<th># of differences</th>
<th>Literal ($S_0$)</th>
<th>Reasoning</th>
<th>Compiled ($S_1$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50%</td>
<td>64%</td>
<td>44%</td>
</tr>
<tr>
<td>2</td>
<td>66%</td>
<td>86%</td>
<td>72%</td>
</tr>
<tr>
<td>3</td>
<td>70%</td>
<td>88%</td>
<td>80%</td>
</tr>
<tr>
<td>4</td>
<td>78%</td>
<td>94%</td>
<td>80%</td>
</tr>
<tr>
<td>Mean</td>
<td>66%</td>
<td>83%</td>
<td>69%</td>
</tr>
</tbody>
</table>

4.4 Final evaluation

Based on the following sections, we keep $k=0$ and use 100 samples to generate predictions. We evaluate on the test set, comparing this reasoning model $S_1$ to two baselines: Literal, an image captioning model trained normally on the abstract scene captions (corresponding to our $L_0$), and Contrastive, a model trained with a soft contrastive objective, and previously used for visual referring expression generation (Mao et al., 2015).

Results are shown in Table 2. Our reasoning model outperforms both the literal baseline and previous work by a substantial margin, achieving an improvement of 17% on all pairs set and 15% on hard pairs.

Figures 4 and 6 show various representative pairs. For comparison, a model with hand-engineered pragmatic behavior—trained using a feature representation with indicators on only those objects that appear in the target image but not the distractor—produces an accuracy of 78% and 69% on all and hard pairs, respectively.
Conclusions

- Standard neural kit of parts for base models
- Probabilistic reasoning for high-level goals
- A little bit of structure goes a long way!
Thank you!
“Compiling” the reasoning model

What if we train the contrastive model on the output of the reasoning model?
Results (dev)

- Literal: 66%
- Compiled: 69%
- Reasoning: 83%