Motivation

Main research question: How can we reliably train language models to generate contextually relevant utterances?

Prior work has investigated training pragmatic language models with communication-based objectives, where neural listeners stand in as communication partners. However…

Challenges include (a) obtaining a well-calibrated listener model, and (b) listener models are domain-specific, which often makes them overconfident about poorly generated utterances [1].

Our work explores whether pragmatic language learning is better with a well-calibrated domain-agnostic listener [2, 3].

Setup

We study the problem of training a pragmatic speaker for reference games with the ShapeWorld dataset [4].

A reference game \((I, t)\) consists of \(n\) images \(I = (i_1, \ldots, i_n)\) and a target image \(t\), with the index \(i\) known only to the speaker.

The objective of the speaker \(f_s\) is to produce an utterance \(u\) which allows the listener \(f_L\) to identify the target \(t\) given the images.

Listeners (L)

We experiment with two types of listeners that differ in which dataset they were trained on. Both listeners are a distribution over possible targets in a reference game. Specifically:

\[
\forall I, u \in \mathcal{I}, t \in \mathcal{D}(I) \quad f_L(I, u) = \exp(g(i)h(n))
\]

where \(g\) and \(h\) are the listener’s image and language encoders, respectively.

Domain-specific (DS) listener \(f_{L,DS}^{DA}\) is trained on the ShapeWorld dataset.

Domain-agnostic (DA) listener \(f_{L,DA}^{DA}\) is the CLIP model (Contrastive Language-Image Pre-training), which is pre-trained on 400 million image (text) pairs collected from the internet [5].

Speakers (S)

Speakers are trained to produce an utterance for the listeners given a game and desired target. Specifically:

\[
f_s(I, u) = p_d(u | g(t), g(i_1, \ldots, g(i_{n-1})))
\]

where \(g\) is the speaker’s image encoder.

Our work considers three base speaker objectives:

- **Domain-agnostic (DA) pragmatic training:**
  \[
  \mathcal{L}_\text{prag}^{DA}(g(I, t), \hat{u} | I, t) = - \log f_{L,DA}(I, \hat{u})
  \]
  - Domain-specific (DS) pragmatic training:
  \[
  \mathcal{L}_\text{prag}^{DS}(g(I, t), \hat{u} | I, t) = - \log f_{L,DS}(I, \hat{u})
  \]
  - Supervised (sup) training:
  \[
  \mathcal{L}_\text{sup}(\hat{u}, u) = - \sum_k \log p_d(u_k | u_k = u_k, I)
  \]

Results and Analysis

**Listener takeaways:**

- Domain-agnostic (DA) listeners are better calibrated than domain-specific (DS) listeners: DA listeners can signal when an utterance is out-of-distribution (OOD).
- However, the DS listener is more confident about in-distribution (ID) utterances!

**Speaker takeaways:**

- Domain specificity and high-confidence in ID utterances is key to training pragmatic speakers: \(\mathcal{L}_\text{sup}\) performs the best. Giving the speaker high rewards when it generates ID utterances is critical.
- Because the DS listener is more confident about ID utterances than the DA listener, the DS listener gives the speaker higher rewards for generating useful ShapeWorld utterances.

Discussion

- We show that the domain specificity of listeners and their high confidence in in-domain utterances is important for training pragmatic speakers.
- Our research can be extended to pragmatic language learning in other domains like COCO [6], where we can experiment with new variations of listener models and speaker objectives.

Examples of generated utterances

<table>
<thead>
<tr>
<th>Objective</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>ground truth</td>
<td>yellow rectangle</td>
</tr>
<tr>
<td>(\mathcal{L}_\text{sup})</td>
<td>si-ten da-da da-da</td>
</tr>
<tr>
<td>(\mathcal{L}_\text{sup})</td>
<td>prize prize prize</td>
</tr>
<tr>
<td>(\mathcal{L}_\text{sup})</td>
<td>yellow rectangle</td>
</tr>
<tr>
<td>(\mathcal{L}<em>\text{sup}) + (\mathcal{L}</em>\text{sup})</td>
<td>religion</td>
</tr>
<tr>
<td>(\mathcal{L}<em>\text{sup}) + (\mathcal{L}</em>\text{sup})</td>
<td>yellow rectangle</td>
</tr>
</tbody>
</table>