Common Ground & Audience Design in Referential Communication

Raquel Fernández
Image description

1. There are several people in chairs and a small child watching one of them play a trumpet.
2. A man is playing a trumpet in front of a little boy.
3. People sitting on a sofa with a man playing an instrument for entertainment.

(example from the visual & linguistic treebank VLT2K dataset)
In image description, several constraints play a role:

1. There are several people in chairs and a small child watching one of them play a trumpet.
2. A man is playing a trumpet in front of a little boy.
3. People sitting on a sofa with a man playing an instrument for entertainment.

*(example from the visual & linguistic treebank VLT2K dataset)*
Image description

1. There are several people in chairs and a small child watching one of them play a trumpet.
2. A man is playing a trumpet in front of a little boy.
3. People sitting on a sofa with a man playing an instrument for entertainment.

*(example from the visual & linguistic treebank VLT2K dataset)*

In image description, several constraints play a role:

- General cooperative principles related to truth and informativeness
Image description

1. There are several people in chairs and a small child watching one of them play a trumpet.
2. A man is playing a trumpet in front of a little boy.
3. People sitting on a sofa with a man playing an instrument for entertainment.

(example from the visual & linguistic treebank VLT2K dataset)

In image description, several constraints play a role:

- General cooperative principles related to truth and informativeness
- Individual constraints: e.g., lexical availability and visual saliency
In image description, several constraints play a role:

- General cooperative principles related to truth and informativeness
- Individual constraints: e.g., lexical availability and visual saliency
- Social interaction constraints: common ground with dialogue partner

1. There are several people in chairs and a small child watching one of them play a trumpet.
2. A man is playing a trumpet in front of a little boy.
3. People sitting on a sofa with a man playing an instrument for entertainment.

*(example from the visual & linguistic treebank VLT2K dataset)*
Re-referring in dialogue

- In dialogue, we often refer to the same entities more than once.
- Besides the constraints mentioned above, subsequent mentions rely on the common ground established with our dialogue partner.
Re-referring in dialogue

- In dialogue, we often refer to the same entities more than once.
- Besides the constraints mentioned above, subsequent mentions rely on the common ground established with our dialogue partner.

A: a white fuzzy dog with a wine glass
B: I see the wine glass dog
A: no I don’t have the wine glass dog
Re-referring in dialogue

- In dialogue, we often refer to the same entities more than once
- Besides the constrains mentioned above, subsequent mentions rely on the common ground established with our dialogue partner

A: a white fuzzy dog with a wine glass
B: I see the wine glass dog
A: no I don’t have the wine glass dog

C: white dog sitting on something red
D: yes, I have the dog on the red chair
C: white dog on the red chair
The PhotoBook dataset

- Two participants see six photos each and need to find out which of three highlighted photos they have in common.
- They can chat freely, without predefined roles.
- The game consists of five rounds, with the set of images changing at every round and some images reappearing.
A: Hi
B: Hello
B: do you have a white cake on multi colored striped cloth?
A: I see a guy taking a picture. What about you?
B: is it of a cake with construction trucks on it?
A: Yeah. I don’t see the cake you mentioned.
A: <common img_2>
A: I see a guy taking a picture. What about you?
B: guy with camera
A: I have the guy with camera
A: the last one is the camera guy
The PhotoBook dataset

Round 2
Participant B

A: I see a guy taking a picture. What about you?

B: *guy with camera*

A: I have the guy with camera

A: the last one is the camera guy
A: I see a guy taking a picture. What about you?
B: guy with camera
A: I have the guy with camera
A: the last one is the camera guy
The PhotoBook dataset

Round 5
Participant A

A: I see a guy taking a picture. What about you?
B: guy with camera
A: I have the guy with camera
A: the last one is the camera guy
The PhotoBook dataset

Co-referring chain: utterances referring to the same target image over a game

A: I see a guy taking a picture. What about you?
B: guy with camera
A: I have the guy with camera
A: the last one is the camera guy
The PhotoBook dataset

Co-referring chain:
utterances referring to the same target image over a game

1. girl on end of bed with computer; she has pigtails
2. Girl with pigtails?
3. Pigtail girl?
4. Pigtails? lol

1. Do you have the girl with the blue umbrella walking by water?
2. I have the girl with the blue umbrella by the water this time
3. What about the blue umbrella girl by the water?
4. Do you have the blue umbrella water girl?

https://dmg-photobook.github.io
2,500 dialogues, 16,525 co-referring chains
Patterns observed in the data

They replicate of previous findings

Patterns observed in the data

They replicate of previous findings


▶ Referring utterances become shorter.
Patterns observed in the data

They replicate of previous findings


- Referring utterances become shorter.
- Increase of content words ratio: more likely to remain.
Patterns observed in the data

They replicate of previous findings


- Referring utterances become shorter.
- Increase of content words ratio: more likely to remain.
- POS distribution: proportion of nouns and adjectives increases.
Patterns observed in the data

They replicate previous findings (Krauss&Weinheimer 1964, Clark&Wilkes-Gibbs 1986, Garrod&Anderson 1987, Clark&Brennan 1991)

- Referring utterances become shorter.
- Increase of content words ratio: more likely to remain.
- POS distribution: proportion of nouns and adjectives increases.
- Sharp decrease of new content words: lexical entrainment.
Patterns observed in the data

They replicate of previous findings


What kind of mechanisms would support these patters?
Patterns observed in the data

They replicate of previous findings


What kind of mechanisms would support these patterns?

Comprehension
  reference resolution

Production
  referring utterance generation
If later references rely on the conversational common ground, they should be more surprising and difficult to **resolve** out of context.
If later references rely on the conversational common ground, they should be more surprising and difficult to resolve out of context.
If later references rely on the conversational common ground, they should be more surprising and difficult to resolve out of context.
$P(w_i | \ldots )$ estimates obtained with GPT-2 fine-tuned on PhotoBook

- Out-of-context surprisal $H(S)$

$$H(S) = - \log_2 P(S) = - \frac{1}{|S|} \sum_{w_i \in S} \log_2 P(w_i | w_1, \ldots, w_{i-1})$$
\[ P(w_i | \ldots) \] estimates obtained with GPT-2 fine-tuned on PhotoBook

- Out-of-context surprisal \( H(S) \)

\[
H(S) = - \log_2 P(S) = - \frac{1}{|S|} \sum_{w_i \in S} \log_2 P(w_i | w_1, \ldots, w_{i-1})
\]

**Language**  
(Giulianelli et al. 2021)

**Language & vision**  
(Takmaz et al. 2022)

Given a referring utterance and the images in the context, CLIP yields softmax probabilities

- Accuracy with highest probability image
- Entropy of the distribution
Language
(Giulianelli et al. 2021)

\[ P(w_i | \ldots) \] estimates obtained with GPT-2 fine-tuned on PhotoBook

- Out-of-context surprisal \( H(S) \)

\[
H(S) = - \log_2 P(S) = - \frac{1}{|S|} \sum_{w_i \in S} \log_2 P(w_i | w_1, \ldots, w_{i-1})
\]

Language & vision
(Takmaz et al. 2022)

Given a referring utterance and the images in the context, CLIP yields softmax probabilities

- Accuracy with highest probability image
- Entropy of the distribution

![Graph showing discriminativeness](chart.png)
Language (Giulianelli et al. 2021)

\[ P(w_i | \ldots ) \] estimates obtained with GPT-2 fine-tuned on PhotoBook

- Out-of-context surprisal \( H(S) \)

\[
H(S) = - \log_2 P(S) = - \frac{1}{|S|} \sum_{w_i \in S} \log_2 P(w_i | w_1, \ldots, w_{i-1})
\]

Language & vision (Takmaz et al. 2022)

Given a referring utterance and the images in the context, CLIP yields softmax probabilities

- Accuracy with highest probability image
- Entropy of the distribution

Higher surprisal and resolution uncertainty in later references without conversational context
\[ H(S) = - \log_2 P(S) = - \frac{1}{|S|} \sum_{w_i \in S} \log_2 P(w_i | w_1, \ldots, w_{i-1}) \]

\[ H(S|C) = - \log_2 P(S|C) = - \frac{1}{|S|} \sum_{w_i \in S} \log_2 P(w_i | w_1, \ldots, w_{i-1}, C) \]
Out-of-context surprisal $H(S)$

In-context surprisal $H(S|C)$

$P(w_i | \ldots)$ estimates obtained with GPT-2 fine-tuned on PhotoBook

Given a referring utterance and the images in the context, CLIP yields softmax probabilities

- Accuracy with highest probability image
- Entropy of the distribution
Language
(Giulianelli et al. 2021)

\[ H(S) = - \log_2 P(S) = - \frac{1}{|S|} \sum_{w_i \in S} \log_2 P(S) \]

Out-of-context surprisal \( H(S) \)

In-context surprisal \( H(S|C) \)

\[ H(S|C) = - \log_2 P(S|C) = - \frac{1}{|S|} \sum_{w_i \in S} \log_2 P(S|C) \]

Given a referring utterance and the images in the context, CLIP yields softmax probabilities

Accuracy with highest probability image

Entropy of the distribution

Uniform low surprisal and resolution uncertainty with conversational context

Language & vision
(Takmaz et al. 2022)
Context dependent generation

- If later references rely on the conversational common ground, context-aware *generation* models will be closer to human patterns
Context dependent generation

- If later references rely on the conversational common ground, context-aware generation models will be closer to human patterns.

- Fine-tuning the model to adapt to the partner (Hawkins et al. CoNLL 2020)
- Relying on episodic memory traces to condition generation (Takmaz et al. EMNLP 2020)
Context dependent generation

- If later references rely on the conversational common ground, context-aware generation models will be closer to human patterns

Takmaz et al (2020): Different encoder-decoder generation models
Context dependent generation

- If later references rely on the conversational common ground, context-aware *generation* models will be closer to human patterns.

Takmaz et al (2020): Different encoder-decoder generation models

- **Ref:** only the visual context, ignoring the linguistic history.
Context dependent generation

- If later references rely on the conversational common ground, context-aware generation models will be closer to human patterns.

Takmaz et al (2020): Different encoder-decoder generation models

- Ref: only the visual context, ignoring the linguistic history.
Context dependent generation

- If later references rely on the conversational common ground, context-aware **generation** models will be closer to human patterns.

Takmaz et al (2020): Different encoder-decoder generation models

- **Ref**: only the visual context, ignoring the linguistic history

- **ReRef**: takes into account both visual and linguistic context, aware of previous mentions.
Similarity to human production patterns
Similarity to human production patterns

**Length in content tokens**
- **First**: Ref > ReRef > Human
- **Subsequent**: Ref > ReRef > Human

**Proportion of nouns**
- **First**: Ref ≈ ReRef > Human
- **Subsequent**: Ref > ReRef > Human

**Content words reused in subsequent mentions**
- Ref: 0%
- ReRef: 20%
- Human: 80%
Similarity to human production patterns

Length in content tokens

<table>
<thead>
<tr>
<th></th>
<th>First</th>
<th>Subsequent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>ReRef</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>Human</td>
<td>8</td>
<td>4</td>
</tr>
</tbody>
</table>

Proportion of nouns

<table>
<thead>
<tr>
<th></th>
<th>First</th>
<th>Subsequent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref</td>
<td>36%</td>
<td>24%</td>
</tr>
<tr>
<td>ReRef</td>
<td>24%</td>
<td>12%</td>
</tr>
<tr>
<td>Human</td>
<td>12%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Content words reused in subsequent mentions

<table>
<thead>
<tr>
<th></th>
<th>Ref</th>
<th>ReRef</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref</td>
<td>0%</td>
<td>20%</td>
<td>40%</td>
</tr>
<tr>
<td>ReRef</td>
<td>20%</td>
<td>40%</td>
<td>60%</td>
</tr>
<tr>
<td>Human</td>
<td>40%</td>
<td>60%</td>
<td>80%</td>
</tr>
</tbody>
</table>

Previous mention: a cake with a Godiva package in the background

- Human: chocolate cake with Godiva package behind it
- ReRef: chocolate cake with Godiva in background
- Ref: do you have a picture of a brown cake on a bed?
Interim summary

In conversation, participants converge on referring expressions that they reuse (“conceptual pacts” become part of the context).

Taking into account this conversational context:

• Makes resolution less effortful, in line with principles of uniform information density and least collaborative effort.
• Helps to generate utterances that are closer to human patterns.
Interim summary

- In conversation, participants converge on referring expressions that they reuse ("conceptual pacts" become part of the context).

- Taking into account this conversational context:
  - Makes resolution less effortful, in line with principles of uniform information density and least collaborative effort.
  - Helps to generate utterances that are closer to human patterns.

- The process whereby participants collaboratively arrive at "conceptual pacts" assumes shared semantic conventions as the starting point for ad hoc shared conventions....
What if the dialogue participants have access to different general conventions and semantic knowledge?

Part 2
Part 2

What if the dialogue participants have access to different general conventions and semantic knowledge?

In other words: How can a cook explain how to make *panna cotta* to someone who has never been in a kitchen?
Part 2

- What if the dialogue participants have access to different general conventions and semantic knowledge?

- In other words: How can a cook explain how to make *panna cotta* to someone who has never been in a kitchen?

- To coordinate not just at the level of dialogue-specific expressions, but also at the level of general semantic knowledge, it is fundamental to be able to represent and reason about others’ mental states.

Data

- We divide the PhotoBook dialogues into 5 domains with minimum vocabulary overlap
Data

We divide the PhotoBook dialogues into 5 domains with minimum vocabulary overlap:

- **Appliances**: A white fridge with the door open
- **Food**: A bowl of raw veggies next to a grapefruit
- **Indoor**: The living room with lamp on a bookshelf
- **Outdoor**: I have a guy doing a handstand on the beach
- **Vehicles**: A parking lot with cars and motorcycle

Example referring utterance per domain
Data

We divide the PhotoBook dialogues into 5 domains with minimum vocabulary overlap

| Domain    | Prop | N    | |V| | Images | Specific | Overlap |
|-----------|------|------|---|---|--------|---------|---------|
| Appliances| 9.4% | 4,310| 1,271 | 36  | 29.5%  | 23.2%   |
| Food      | 12.4%| 5,682| 1,646 | 36  | 43.3%  | 22.9%   |
| Indoor    | 26.4%| 12,088| 2,477 | 96  | 44.3%  | 26.0%   |
| Outdoor   | 35.9%| 16,427| 2,858 | 108 | 47.0%  | 26.2%   |
| Vehicles  | 15.8%| 7,234| 1,738 | 48  | 36.0%  | 26.2%   |
| All       | 100% | 45,741| 6,038 | 324 | -      | -       |
The speaker
Visually conditioned language model

- **Input:** visual context including target image
- **Goal:** generate a referring utterance for the target
- **Training:** trained on all domains — “proficient speaker”
The listener

Discriminator model

- **Input:** visual context and utterance
- **Goal:** identify the target image the utterance refers to
- **Training:** trained on a single domain — “domain-specific listener”
Resolution performance without adaptation

With domain-specific listeners, if the speaker does not adapt then communication is unsuccessful:

- High accuracy for in-domain settings (diagonal)
- Near chance accuracy (16%) for out-of-domain cases

<table>
<thead>
<tr>
<th>Input image domain</th>
<th>app</th>
<th>food</th>
<th>indoor</th>
<th>outdoor</th>
<th>vehi</th>
</tr>
</thead>
<tbody>
<tr>
<td>appliances</td>
<td>57.61</td>
<td>20.10</td>
<td>19.92</td>
<td>21.27</td>
<td>15.98</td>
</tr>
<tr>
<td>food</td>
<td>19.11</td>
<td>54.29</td>
<td>18.60</td>
<td>18.85</td>
<td>18.85</td>
</tr>
<tr>
<td>indoor</td>
<td>22.71</td>
<td>19.65</td>
<td>53.62</td>
<td>20.82</td>
<td>16.77</td>
</tr>
<tr>
<td>outdoor</td>
<td>15.08</td>
<td>21.46</td>
<td>19.62</td>
<td>52.93</td>
<td>17.69</td>
</tr>
<tr>
<td>vehicles</td>
<td>16.36</td>
<td>16.17</td>
<td>17.41</td>
<td>20.13</td>
<td>43.08</td>
</tr>
</tbody>
</table>
Plug-and-Play Theory of Mind

- How can the speaker adapt its utterances to the listener’s knowledge?
- On the fly, without fine-tuning the language model permanently?
Plug-and-Play Theory of Mind

- How can the speaker adapt its utterances to the listener’s knowledge?
- On the fly, without fine-tuning the language model permanently?
- Inspired by work on controlled text generation (Dathathri et al., 2020), we explore a “plug-and-play” approach

(Dathathri et al., 2020)
Plug-and-Play Theory of Mind

- Knowledge asymmetry
- The speaker tailors its utterance about a food image for a listener who does not know about food
- The speaker’s simulator module guides this adaptation via self-monitoring loop
Plug-and-Play Theory of Mind

- The simulator is trained to predict the behaviour of a domain-specific listener, given a “planned” utterance and visual context
  - Simplification: the speaker knows the type of listener a priori
Plug-and-Play Theory of Mind

- The simulator is trained to predict the behaviour of a domain-specific listener, given a “planned” utterance and visual context
  
  - Simplification: *the speaker knows the type of listener a priori*

- It refines the speaker’s utterance plan iteratively (∼ self-monitoring) until it considers it sufficiently discriminative for the listener
  
  - “*Would the listener be able to resolve this utterance?*” If the prediction is negative, this triggers an update to the speaker’s decoder initial state, and the utterance gets updated
Plug-and-Play Theory of Mind

- The simulator is trained to predict the behaviour of a domain-specific listener, given a “planned” utterance and visual context
  - Simplification: *the speaker knows the type of listener a priori*

- It refines the speaker’s utterance plan iteratively (≈ self-monitoring) until it considers it sufficiently discriminative for the listener
  - “*Would the listener be able to resolve this utterance?*” If the prediction is negative, this triggers an update to the speaker’s decoder initial state, and the utterance gets updated

- Finally, the referring utterance is overtly passed on to the listener (who may or may not be able to resolve it — the simulator is not perfect!)
Does it work?

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Audience-aware</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OOD</strong></td>
<td>19.06 ± 0.47</td>
<td>26.74 ± 1.48</td>
</tr>
<tr>
<td><strong>IND</strong></td>
<td>52.30 ± 1.10</td>
<td>71.77 ± 2.16</td>
</tr>
</tbody>
</table>

averages across domains

- Audience-aware adaptation leads to significant increases in accuracy
- Including more than 7% in scenarios where the image domain is not known to the listener (OOD)
Does it work?

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Audience-aware</th>
</tr>
</thead>
<tbody>
<tr>
<td>OOD</td>
<td>19.06 ± 0.47</td>
<td>26.74 ± 1.48</td>
</tr>
<tr>
<td>IND</td>
<td>52.30 ± 1.10</td>
<td>71.77 ± 2.16</td>
</tr>
</tbody>
</table>

*averages across domains*

- Audience-aware adaptation leads to significant increases in accuracy
- Including more than 7% in scenarios where the image domain is not known to the listener (OOD)

How does it work?
Qualitative examples

PhotoBook participant:  *I have the pink food truck again ... white shirt lady*

Generated not adapted:  *girl at black phone, red truck, brown hair, pink*

Generated adapted:  *pink donuts*
Qualitative examples

PhotoBook participant: green salad with a person holding up a portion with fork?

Generated not adapted: I have one more maybe round you think that has a lime green shaped greens, a salad?

Generated adapted: must bookshelves in the salad?
Probing for domain adaptation

- We expect $h_0$ to carry information about the **target image domain**, because it is the result of encoding such image.

- Indeed, using a diagnostic classifier we can predict the image domain from the $h_0$ with 100% accuracy.
Probing for domain adaptation

- We expect $h_0$ to carry information about the **target image domain**, because it is the result of encoding such image.

- Indeed, using a diagnostic classifier we can predict the image domain from the $h_0$ with 100% accuracy.

- Does $h_0$ carry information about the **listener’s domain**? Not before adaptation.
We expect $h_0$ to carry information about the target image domain, because it is the result of encoding such image.

Indeed, using a diagnostic classifier we can predict the image domain from the $h_0$ with 100% accuracy.

Does $h_0$ carry information about the listener’s domain? Not before adaptation.

With adaptation, the encoding of the image domain deteriorates, while the listener’s domain becomes highly predictably.
Properties of the adapted utterances

- Adapted utterances: when the speaker *believed* the utterance would lead to communicative success
Properties of the adapted utterances

Adapted utterances: when the speaker believed the utterance would lead to communicative success

More successful adapted utterances contain:

- words with lower age of acquisition
- lower rate of lexical choice from the target image vocabulary and higher rate of words from the listener vocabulary
Summing Up

- We decide what to say and how to say it on the basis of what we share with our dialogue partner.

- It is an open question how such accommodation can be modelled in computational agents.
Summing Up

- We decide what to say and how to say it on the basis of what we share with our dialogue partner.

- It is an open question how such accommodation can be modelled in computational agents.

- This talk: A few proposals for analysing context dependence of repeated references, generating dialogue-aware references, adapting utterances via theory of mind simulation.
Summing Up

- We decide what to say and how to say it on the basis of what we share with our dialogue partner.

- It is an open question how such accommodation can be modelled in computational agents.

- This talk: A few proposals for analysing context dependence of repeated references, generating dialogue-aware references, adapting utterances via theory of mind simulation.

- Many limitations remain: for example, our agent models (speaker, simulator, listener) are pertained and remain frozen. This has advantages but dramatically oversimplifies the dynamics of interaction.
Thanks

Niccolò Brandizzi  Ece Takmaz  Mario Giulianelli  Sandro Pezzelle  Arabella Sinclair  Janosch Haber

• Takmaz et al. Refer, reuse, reduce: Generating subsequent references in visual and conversational contexts. EMNLP 2020.
• Giulianelli et al. *Is information density uniform in task-oriented dialogues?* EMNLP 2021.