Today

• Introduction: Dialogue and Grounded Dialogue
• Grounded Dialogue Datasets
• A Bit of Theory
• Case Study: Reference Games
• Outlook: IR to Model Human Behaviour?
Dialogue

*What is it and why do we care?*

- **What?** Using language for inter-personal communication and interaction
- **Why?** The primary form of language use and language learning
- **Where?** Face-to-face, on the phone, on Zoom, on Signal, on Reddit, ...
- **How?** Linguistics, psychology, sociology, cognitive science, mathematics, ...
Grounded Dialogue

*Interactive language use in context*

Dialogue (just like any type of language use) happens in **context**, in an **environment**. Speakers communicate to change the state of the environment and achieve **goals**. Communicating is an **action** — through dialogue, an **interaction**.
Grounded Dialogue Modelling

*The study of interactive language use in context*

Dialogue (just like any type of language use) happens in context, in an environment. Speakers communicate to change the state of the environment and achieve goals. Communicating is an action — through dialogue, an interaction.

- What is the relevant context of an interaction?
- How does the context relate to a speaker’s communicative goals?
- What are the decision making strategies that humans follow to choose words and achieve goals in their environment?
- Can we replicate them in a computer system?
Grounded Dialogue Modelling

The study of interactive language use in context

Dialogue (just like any type of language use) happens in context, in an environment. Speakers communicate to change the state of the environment and achieve goals. Communicating is an action — through dialogue, an interaction.

- What is the relevant context of an interaction?
- How does the context relate to a speaker’s communicative goals?
- What are the decision making strategies that humans follow to choose words and achieve goals in their environment?
- Can we replicate them in a computer system?
Grounded Dialogue Datasets
Map Task


Spoken dialogues (transcribed): instruction giving and following to navigate to a point on a map.

- **g** start off above the diamond mine
- **f** okay yeah
- **g** now go south from the diamond mine until you are just above the desert
- **f** so that's with the diamond mine on on your right
- **g** that's that's correct uh-huh
- **g** and go below the diamond mine
- **f** mmhmm
- **g** and below the graveyard below the graveyard but above the carved wooden pole
- **f** oh hang on i don't have a graveyard
PhotoBook

Haber et al., 2019. ACL.

Written cooperative reference game: describe images in turn to find common sets of photographs.

Link to: PhotoBook website (data, visualisation, code)
PersonaChat

Zhang et al., 2018. ACL.

Written *chit-chat* dialogue: given a character description, chat with another person naturally and try to get to know each other.

Links to:
GitHub readme
ParlAI website

---

**Persona**

I bought my first home.
I love to barbecue.
I live in Springfield.
I’m a writer.

**Persona**

I weight 300 pounds.
I am not healthy.
I am a man.
I like The Godfather.

---

Hello how are you, I am new to the Springfield area.

Hi! Seen any good movies lately?

I have been to the movies.

I love The Godfather, one of my favorites! Was that filmed?

I don’t believe so. I don’t watch movies more of a writer.

More (Grounded) Dialogue Datasets

A few useful resources

- https://parl.ai/docs/tasks.html
- https://docs.google.com/spreadsheets/d/1N5_5gBKlGR-OrigRNct4jQ6iEqSycyqcoN61JpsHFDQ/htmlview
A Bit of Theory
The ‘environment’ in which the interaction takes place

$w \in W$

Speaker + Audience

Context
Communicative Goal

*change of the state of the environment*

For communication to be successful, the audience must be able to reconstruct the speaker's communicative goal.

Communicative goals shape and constrain the speaker's production choices: different types of utterance correspond to different goals.
Costs

*The cognitive and physical efforts required to communicate*

**Production costs**
Speaker executes a bit of behaviour (e.g., speaking or typing) meant to be perceived by the audience in order to convey their communicative intent.

**Comprehension costs**
Audience attends to and processes the behaviour executed by the speaker in order to reconstruct the speaker’s communicative intent.

*Speakers estimate these costs and take them into account when they choose words.*
Utility

*The cognitive, physical, and social effects of a communication act*

Inversely proportional to:
- the joint production and comprehension costs (*collaborative effort*)
- the distance between the new and the intended state of the world

Directly proportional to:
- the positive cognitive, physical, and social effects derived from achieving the intended new state (the communicative goal)
Grounded Dialogue Modelling

The study of interactive language use in context

- What is the relevant context of an interaction?
- How does the context relate to a speaker’s communicative goals?
- What are the decision making strategies that humans follow to choose words and achieve goals in their environment?
- Can we replicate them in a computer system?
Case Study: Reference Games
PhotoBook

Haber et al., 2019. ACL.

Written cooperative reference game: describe images in turn to find common sets of photographs.
Reference chains

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?
The ‘environment’ in which the interaction takes place

Visual context

Speaker A + Speaker B

Conversational context
Context

The ‘environment’ in which the interaction takes place

Visual context

Reference chain

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?
Communicative Goal

A change of the state of the environment
Costs

*The cognitive and physical efforts required to communicate*

**Production costs**
Utterance planning, typing, editing, ...

**Comprehension costs**
Reading, interpretation / reference resolution
Noisy Channel Model of Communication

Claude Shannon, 1948

SPEAKER

intent → signal

ADDRESSEE

received signal → interpretation

noisy channel
Noisy Channel Model of Communication

Claude Shannon, 1948
Noisy Channel Model of Communication

Claude Shannon, 1948

SPEAKER

intent → signal → received signal

ADDRESSEE

interpretation

noisy channel

realisation effort → processing effort

33
Computational Estimates of Processing Effort

*via Shannon information content:*  
$$-\log P(X)$$

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

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

$P(w_i \mid \ldots)$ estimates obtained with **GPT-2** (Radford et al., 2018), a neural language model which we fine-tune on PhotoBook.
Results: PhotoBook Reference Chains

Speakers reduce collaborative effort

- Reduction of processing effort
- (Information compression)
- Reduction of realisation effort
Noisy Channel Model of Communication

Claude Shannon, 1948

SPEAKER

intent → signal → noisy channel

ADDRESSEE

received signal → interpretation

reference resolution effort
Computational Estimates of Resolution Effort

How well does the utterance describe the target image?

Descriptiveness: $CLIPScore(\text{image, utterance})$

High descriptiveness = predictable image-utterance matching

Estimates obtained with CLIP (Radford et al., 2021) a neural vision & language model
(Contrastive Language-Image Pre-training via symmetric image-text matching loss)
Utility

The cognitive, physical, and social effects of a communication act
Computational Estimates of Utility

**Positive utility: Is the communicative goal achieved?**

Discriminativeness: task success (accuracy)
1 if target has the highest probability, otherwise 0

Estimates obtained with **CLIP** (Radford et al., 2021) a neural vision & language model
(Contrastive Language-Image Pre-training via symmetric image-text matching loss)
Results: PhotoBook Reference Chains

Speakers reduce collaborative effort while ensuring task success

Descriptiveness decreases over time yet discriminativeness is not significantly affected.
Grounded Dialogue Modelling

The study of interactive language use in context

๏ What is the relevant context of an interaction?
๏ How does the context relate to a speaker’s communicative goals?
๏ What are the decision making strategies that humans follow to choose words and achieve goals in their environment?
๏ Can we replicate them in a computer system?
Grounded Dialogue Modelling

The study of interactive language use in context

๏ What is the relevant context of an interaction?
๏ How does the context relate to a speaker’s communicative goals?
๏ What are the decision making strategies that humans follow to choose words and achieve goals in their environment?
๏ Can we replicate them in a computer system?

Refer, Reuse, Reduce
Generating Subsequent References in Visual and Conversational Contexts

Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, Raquel Fernández
Institute for Logic, Language and Computation
University of Amsterdam
{e.takmaz|m.giulianelli@s.pezzelle|a.j.sinclair|raquel.fernandez}@uva.nl

Abstract
Dialogue participants often refer to entities or situations repeatedly within a conversation, which contributes to its cohesiveness. Subsequent references exploit the common ground accumulated by the interlocutors and hence have several interesting properties, namely, they tend to be shorter and reuse expressions that were effective in previous mentions. In this paper, we tackle the generation of first and subsequent references in visually grounded dialogue. We propose a generation model that produces referring utterances grounded in both the visual and the conversational context. To
Dialogue Modelling Group

Research conducted in our research group at the ILLC with:

Ece Takmaz
Sandro Pezzelle
Arabella Sinclair
Raquel Fernández

More details & results on other corpora:


• E. Takmaz, S. Pezzelle, R. Fernández. Less descriptive yet discriminative: Quantifying the properties of multimodal referring utterances via CLIP. CMCL Workshop, ACL 2022.
Outlook:  
IR to Model Human Behaviour?
Outlook

• Computational modelling of human production strategies using pre-trained language & multimodal models.

• Reference resolution in visual contexts is essentially a retrieval task.

• This idea can be extended to other modes of productions: e.g. summarisation, translation, text simplification, ...
Thanks

For any questions: m.giulanelli@uva.nl