Disambiguating Reference in Visually Grounded Dialogues through Joint Modeling of Textual and Multimodal Semantic Structures

Shun Inadumi [1,2], Nobuhiro Ueda [3], Koichiro Yoshino [4,2,1]
1. NAIST, 2. RIKEN GRP, 3. Kyoto University, 4. Science Tokyo

Understanding References in Dialogues

TRR: Textual Reference Resolution

Identifies **textual reference** relations **between phrases**

TRR partially consists of the following tasks:

- **Coreference** Resolution — this = the coffee cup
- **Predicate-argument sturucture** analysis — take →ACC the coffee cup

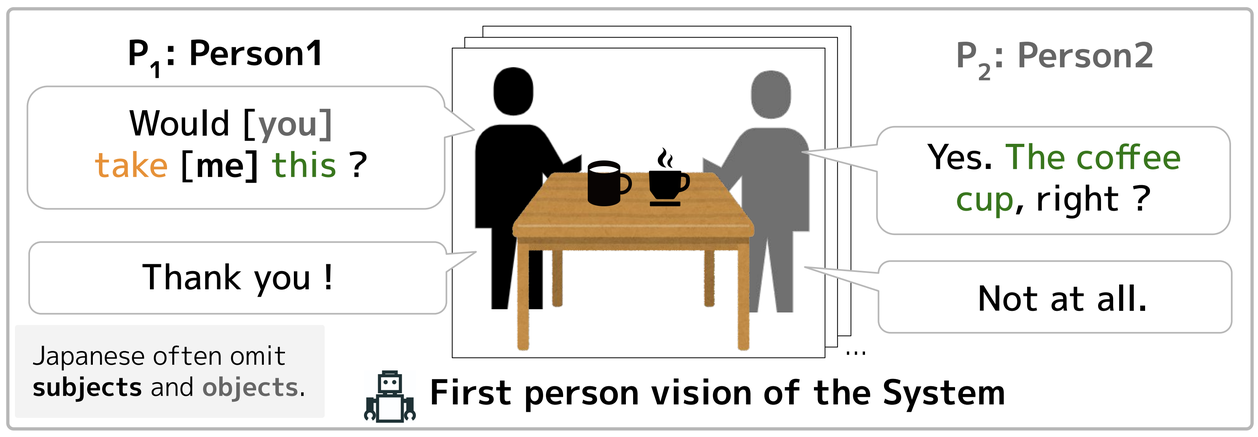
MRR: Multimodal Reference Resolution [Ueda+, 2024]

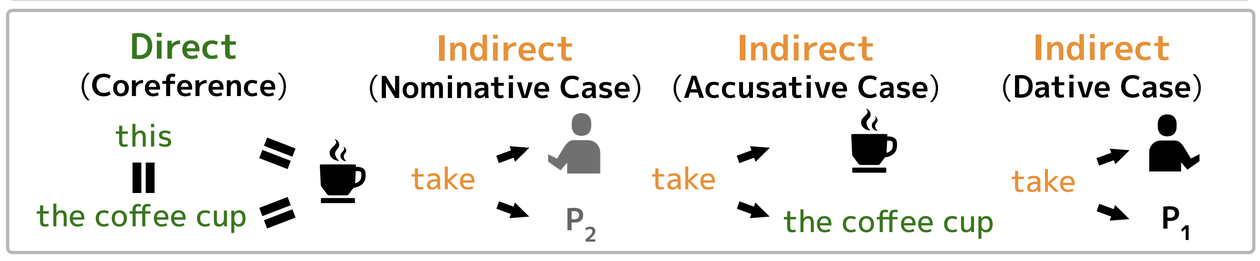
Identifies **phrase-to-objects direct/indirect references**

- Phrase Grounding [Plummer+, 2015]: task limited to direct references. — the coffee cup =

MRR enables systems to understand **dialogue events** — such as "**who does what to whom**"— linked to objects.

Example of the system analyzes a two-person dialogue

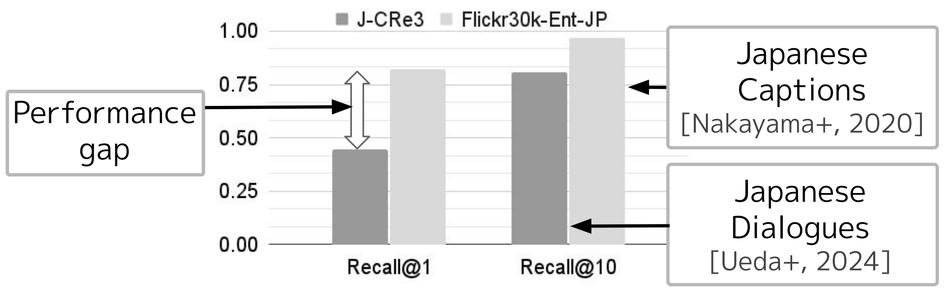




Limitations of Existing Models

- GLIP [Li+, CVPR2022]: A phrase grounding model trained on large-scale **image-caption pairs** with direct references.
- In dialogue parsing, GLIP struggles with:

1. Resolving direct references made via **pronouns** — this =
2. Parsing indirect references involving **ellipses** — take →DAT,NOM ,



Phrase grounding results of GLIP

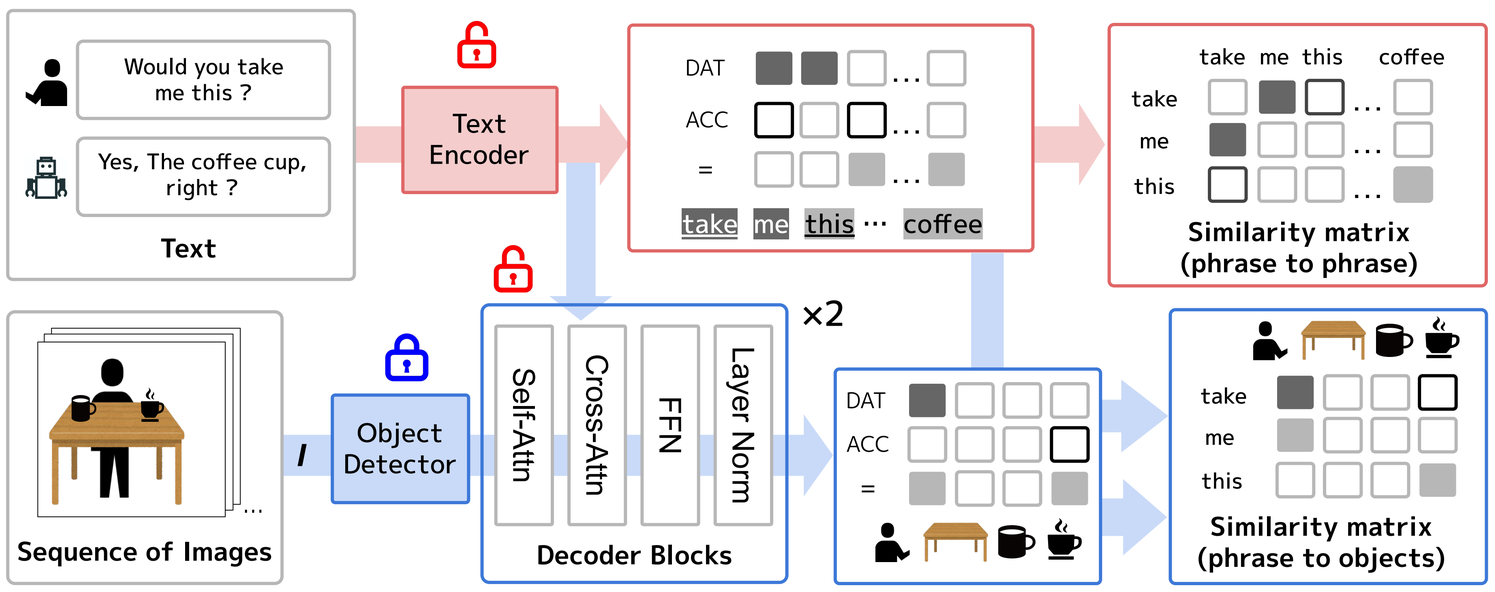
By resolving ambiguities in 1 and 2 , we aim to better understand real-world dialogues.

Proposed Framework

- By incorporating **textual references**, we can improve MRR performance.

e.g.) If the coffee cup — this is known, this = can be uniquely identified.

- We propose a framework to jointly model **TRR** and **MRR**.



Phrase Grounding Results

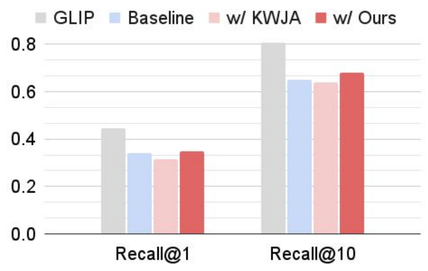
Compared models:

- Baseline
- Baseline w/ Ours
- Baseline w/ KWJA [Ueda+, 2023]
- GLIP [Li+, CVPR2022]

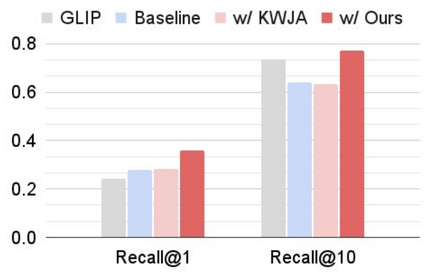
Phrase grounding model with coreference resolution (fine-tuned on Japanese data)

Pre-trained on English data [Krishna+, 2017, Hudson+, 2019]

Main Results:

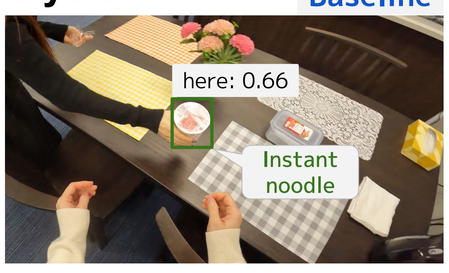


Japanese Dialogue (Overall, 996)



Japanese Dialogue (**Pronouns**, 120/996)

Analysis:



Can you put [the water] in here since it comes up ?

Japanese omits [ ] phrases.

Improvements from coreference resolution:

- **Improved pronoun phrase grounding**
- **Increased confidence scores for pronouns**

MRR Results

Parsing Targets:

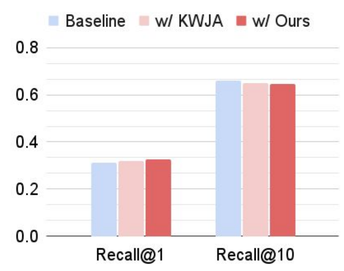
- Direct Reference
- Indirect References:
  - Predicate-argument structures (PAS)
  - Bridging Anaphora

Compared models:

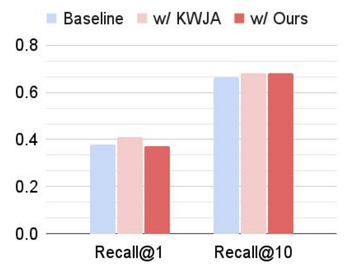
- Baseline
- Baseline w/ Ours
- Baseline w/ KWJA

MRR models with TRR

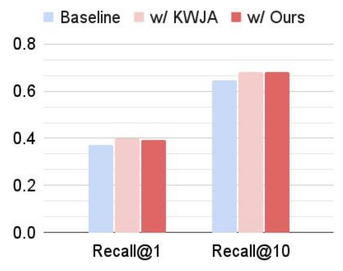
Main Results:



Direct

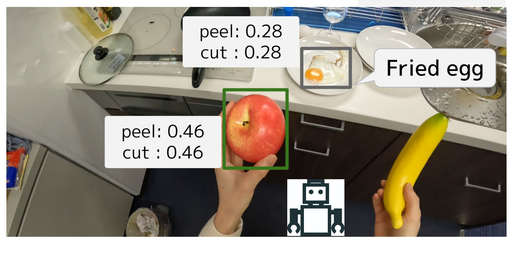


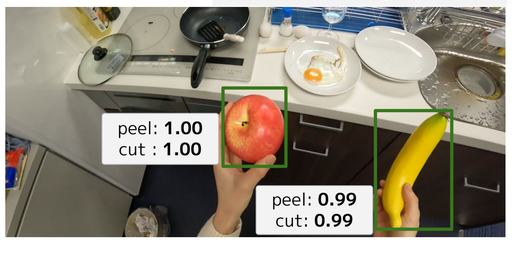
Indirect (PAS)



Indirect (Bridging)

Analysis:





Shall we peel both [the apple and the banana]? Then, let's cut [them] into portions for three people.

Japanese omits [ ] phrases.

Improvements from TRR:

- **Improved indirect reference performance**
- **Increased confidence scores for predicates**