

Addressing Referential Ambiguity in Visually Grounded Dialogue

氏 名: 稲積 駿

研 究 室 名: ソーシャルコンピューティング研究室

主指導教員名: 荒牧 英治

内容梗概:

Human communication is ambiguous. In everyday interaction, people do not always express their intentions explicitly; instead, they rely on contextual information, such as the shared environment and shared knowledge, to interpret utterances efficiently. Such context is naturally acquired, interpreted, and used through human interaction. However, enabling intelligent systems to use context in this way remains challenging, especially when they interact with their users in real-world environments.

This dissertation addresses this challenge by studying how systems can use context to handle ambiguity. In particular, we focus on referential ambiguity in visually grounded dialogue, where the utterance alone is insufficient to determine which entity a context-dependent linguistic expression refers to in a shared visual environment.

To this end, this dissertation examines how referential ambiguity can be handled using both linguistic context, such as dialogue history, and non-linguistic context, including visual information and eye-gaze information from users.

We study two complementary problems. The first is ambiguity resolution, where a system uses context to determine a unique interpretation when the available context is sufficient. The second is ambiguity recognition, where a system estimates multiple plausible interpretations when the available context is insufficient to determine a unique one.

This dissertation pursues these problems across vision-and-language tasks that connect linguistic and visual information. First, we introduce a dataset and develop a model that uses eye-gaze information to answer ambiguous visual questions. Second, we propose a framework for resolving ambiguous references in visually grounded dialogue by using linguistic context. Finally, we introduce a benchmark for evaluating whether vision-language models can recognize unresolved ambiguity by estimating plausible candidate objects.

Taken together, these studies show that systems can use linguistic and non-linguistic context to resolve referential ambiguity arising from pronouns and ellipsis, while also highlighting the difficulty of recognizing unresolved ambiguity. The contributions provide a fundamental step toward natural communication and collaboration between intelligent systems and their users in real-world environments.