# A Study on Advanced Common Grounding in Natural Language Dialogue Systems
# （自然言語対話システムにおける発展的な基盤化の研究）

by

Takuma Udagawa
宇田川拓真

A Doctor Thesis
博士論文

Submitted to
the Graduate School of the University of Tokyo
on June 4, 2021
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Information Science and Technology
in Computer Science

Thesis Supervisor: Akiko Aizawa　相澤彰子
Professor of Computer Science

**ABSTRACT**

One ultimate goal of natural language processing is to develop systems that can understand and communicate reliably in natural language. Towards this end, we focus on the fundamental aspect of communication called common grounding, which refers to the collaborative process of creating and maintaining mutual understandings (i.e. common ground). Despite the long history of study on this topic, we raise three major limitations in existing research. First, existing works mostly focus on restricted domains and task settings where advanced common grounding is not required. Secondly, how to evaluate and analyze common grounding remains an open problem. Finally, existing dialogue systems still have limited capability of common grounding: to be specific, traditional (pipeline) dialogue systems lack the flexibility and end-to-end (fully data-driven) systems lack the robustness required for advanced common grounding.

To address these problems, we propose a novel research platform to study advanced common grounding in natural language communication. In Chapter 1, we introduce the precise scope of this thesis, namely the viewpoint of common grounding as entity-level alignment. In Chapter 2, we give an overview of the existing literature on common grounding, dialogue systems and symbol grounding research. In Chapter 3, we formally define the collaborative reference task to measure the ability of creating common ground. To study advanced common grounding, we propose a novel task setting under continuous and partially-observable context, where collaborative resolution of complex ambiguity, uncertainty and misunderstandings is required. Based on this task design, we developed OneCommon Corpus: a large-scale dataset of 6,760 dialogues for conducting empirical studies. In Chapter 4, we propose an annotation framework of reference resolution to interpret common grounding across modalities. We show that our approach is reliable, scalable, and useful for investigating the strategies of both humans and imperfect dialogue systems. In Chapter 5, we provide additional annotation of spatial expressions on top of reference resolution. Based on these annotations, we capture fine-grained linguistic structures as well as how they are externally grounded at the level of symbol grounding. In Chapter 6, we generalize our collaborative reference task under dynamic context to study advanced skills of maintaining common ground. Based on this task formulation, we developed Dynamic-OneCommon Corpus: another large-scale dataset containing 5,617 dialogues. Through our empirical studies, we demonstrate advanced strategies required under this setting, such as the usage of complex spatio-temporal expressions and previous common ground to retain or update common ground.

Based on this platform, we enable various dialogue systems to be evaluated, analyzed and improved in terms of advanced common grounding. In Chapter 7, we explore the future prospects of our research, including further ideas on the task design, model improvements, and real-world applications. In Chapter 8, we summarize the contributions of this thesis. Overall, we expect our study to be a fundamental step towards developing dialogue systems that can collaborate with humans reliably in natural language.

## 論文要旨

自然言語処理研究における究極的な目標の一つは、人間のように言語を理解し、信頼性の高いコミュニケーションを行える対話システムの実現である。この目標に向けて、本論文ではコミュニケーションの根幹を成す基盤化という現象に着目する。基盤化とは、共通理解（共通基盤）を形成・維持する一連のプロセスを指し、重要な研究対象として長い歴史を持つ。しかし、既存研究には大きく三つの問題点がある。第一に、既存の対話研究はドメイン・タスク設計の観点から限られており、高度な基盤化を要請しない。第二に、基盤化の能力の評価・分析を行うことは一般に難しい。第三に、既存の対話システムの基盤化能力の限界がある。具体的には、従来の（パイプライン型の）対話システムは柔軟性に欠けており、（完全にデータ駆動型の）end-to-end 対話システムは頑健性に欠けているため、どちらも複雑な状況に対処する高度な基盤化を行うことができない。

以上の問題点を解決するために、本論文では発展的な基盤化の研究を推進するための新たな研究プラットフォームを提案する。1 章では基盤化をエンティティレベルのアラインメントとして捉えることで、本研究の対象領域の明確な導入を行う。2 章では重要な関連研究、特に基盤化・対話システム・記号接地問題の分野を概観する。3 章では正確な共通基盤の形成を評価するための協調的参照タスクを定義する。その上で、連続的な情報および部分観測的な状況を扱うタスク設定を導入することで、曖昧性・不確実性・誤解の解消を含む高度な基盤化を要請する。このタスク設定に基づき、6,760 対話を含む大規模な OneCommon コーパスを構築し、様々な実証的研究を可能にする。4 章では参照解決のアノテーションに基づくモダリティ横断的な基盤化の解釈手法を提案する。提案手法によって分析の信頼性と拡張性を担保しつつ、人間だけでなく不完全な対話システムの解釈が可能になることを示す。5 章では参照解決に加えて空間表現のアノテーションを組み合わせたさらなる分析を行う。これによって対話中のより詳細な言語構造を捉えつつ、それらのモダリティ横断的な（記号接地問題のレベルでの）深い分析が可能になることを示す。6 章では協調的参照タスクを時系列的に拡張することで、動的な環境における共通基盤の維持を評価する新たなタスクの設計を行う。このタスク設計に基づき 5,617 対話を含む Dynamic-OneCommon コーパスを新たに構築し、動的な環境では複雑な時空間表現や過去の共通基盤を利用した高度な基盤化が要請されることを示す。

本論文の提案プラットフォームを用いることで、様々な対話システムを発展的な基盤化の観点から実装・評価・分析することが可能になる。7 章では更なるタスク設計のアイデア、モデルの改善手法、現実問題への応用を含めた今後の研究の展望について考察する。最後に 8 章では本研究全体を総括する。以上の貢献により、本論文は自然言語を通じて人間と協調できる、信頼性の高い対話システムの実現に向けた重要な基礎を構築する。

# Acknowledgements

First and foremost, I would like to express my sincere gratitude to Professor Akiko Aizawa for her invaluable guidance. Throughout the writing of this PhD thesis, I've been always encouraged by her unwavering belief, understanding and support in my research. It was truly a life-changing experience to spend the 5 years under her supervision. I will cherish everything I've learned from her in my future career, including the virtues of modesty, kindness and respect for others.

I must also express my gratitude to Asst. Prof. Saku Sugawara for all the inspiring discussions. His pivotal insight on natural language understanding research had a profound influence on my studies. I'm also deeply grateful for always providing me constructive comments on my research plans and paper writing.

I would like to extend my sincere thanks to the members of our laboratory, especially Vitou Phy and Takato Yamazaki for willingly participating in my annotation projects. I owe my special thanks to the secretory Noriko Katsu for assisting many aspects of my laboratory life. I was fortunate to be able to work and spend the time together with all the great members at National Institute of Informatics.

I am also grateful to the members of my PhD thesis commitee, Assoc. Prof. Hideki Nakayama, Prof. Imari Sato, Prof. Masami Hagiya, Prof. Ryuichiro Higashinaka, and Prof. Yusuke Miyao. They provided me important feedback on improving the early draft and presentation of this thesis.

Finally, my deepest appreciation goes to my family: especially my parents, siblings, nephew, and partner. Without their tremendous understanding and support, this milestone could not have been achieved. I dedicate this thesis to them.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 The Common Grounding Problem

Human communication is extraordinary. Unlike other primates, we use *symbolic* natural language which can be assembled grammatically to express countless meanings, from abstract to concrete (Tomasello, 2009). Natural language is so versatile that most (if not all) of our knowledge is expressed, shared and elaborated through this medium. Hence, it is not surprising that such linguistic competence has been considered as a hallmark of human-level intelligence (Turing, 1950).

But what makes human communication so reliable? At the heart of this question lies the problem of *common grounding*. Common grounding is the process of creating and maintaining mutual understandings (i.e. *common ground*), which is a critical aspect of sophisticated human communication. It is only through this collaborative effort that we can ensure the reliability of our shared understandings (Clark, 1996).

To illustrate how this process unfolds in natural language dialogues, we first introduce the *contribution theory* of Clark and Schaefer (1989), which remains influential to date. According to this theory, information in dialogue transitions through 2 phases: the *presentation phase*, where it gets first introduced by the speaker, followed by the *acceptance phase*, where it gets acknowledged by the listener(s). It is only after the positive feedback from the listener(s) in the acceptance phase that the information is added (or *contributed*) to their common ground.[1]

To exemplify this process, we show an actual conversation reported in Sacks et al. (1974).

(1)
a. A: Uh you been down here before haven't you.
b. B: Yeah.
c. A: Where the sidewalk is?
d. B: Yeah,
e. A: Where it ends,
f. B: Goes all the way up there?
g. A: They come up to there, yeah.

In the first utterance (a), speaker A inquires whether B has been to the place ("*here*") before, which B recognizes and replies with a positive answer (b). These two utterances form a contribution, and the inquired fact becomes mutually accepted among A and B. However, there remains a potential ambiguity in the deictic reference "here", so the conversation continues to resolve it through clarifications and elaborations (c-g). Based

[1] Note that a piece of information can be presented in various ways, including direct assertions as well as indirect *presuppositions*, e.g. implicitly assumed in a question (Stalnaker, 1978). Similarly, information can be accepted directly through acknowledgements (like "yes", "okay") or more indirectly, e.g. through the initiation of the next relevant contribution (Cho and May, 2020).

on this process, the speakers can ensure the detail and reliability of common ground, and if there were any misunderstandings, they can be *repaired* through correction (e.g. "No, it's the *opposite* side of the sidewalk.") (Schegloff et al., 1977).

Through the accumulation of this process, humans can develop a substantial amount of common ground foundational in various aspects of our daily life. Without the ability of reliable common grounding, the productive, stable and efficient collaboration in complex human society is unimaginable to be achieved.

## 1.2 Limitations of Existing Research

The previous discussion corroborates the importance of common grounding in human communication. But how is this problem being addressed in the fields of artificial intelligence (AI) and natural language processing (NLP), and what are the main challenges that hinder its progress? Despite the long history of study on this topic, we raise three major limitations of existing research.

### Limitation on the Task Settings

In dialogue research, *task design* is an important factor that determines the subject of study (e.g. the complexity and requisite strategy of common grounding). However, existing dialogue tasks mostly focus on restricted domains and task settings where advanced, full-fledged common grounding is not required. To be specific, most existing tasks only need to deal with the following types of information which make the requisite skills of common grounding relatively trivial.

- *Categorical* information: Traditionally, the type of information that dialogue systems handle has been limited to categorical/discrete information. For instance, the information of color would be discretized into predefined categories (e.g. "red", "blue", "green", and so forth) so that they can be treated with structured databases and frame-based dialogue state tracking (Henderson, 2015).

  However, this setting ignores the aspect of many concepts being gradable and unbounded (Lakoff, 1987; Paradis, 2008) and introduces minimal ambiguity or uncertainty to be dealt with symbolic natural language. Consequently, they require minimal effort of semantic coordination (e.g. disambiguations and clarifications) in the process of common grounding.

- *Fully-observable* information: It is also common in prior works to assume that the information is fully-observable, i.e. all agents have the same, complete information of the environment. For instance, agents often have identical observations (e.g. the same set of images) to discuss in the dialogue (Zarrieß et al., 2016; De Vries et al., 2017; Shore et al., 2018).

  However, this assumption is unrealistic in many situations where we only have partial, private information of the environment. Due to the lack of such information asymmetry, existing settings introduce minimal misunderstandings or partial understandings that need to be resolved through common grounding.

- *Static* information: Finally, existing dialogue tasks mostly focus on static information and environments that do not change over time. For instance, dialogue contexts like databases (He et al., 2017), images (Haber et al., 2019) and embodied environments (de Vries et al., 2018; Thomason et al., 2019) are usually assumed to be fixed/stationary in the course of the dialogue.

However, real-world environments are *dynamic* and require continuous effort of maintaining common ground. In the static settings of existing works, there are minimal requirements for updating common ground and adapting to the evolving situations, e.g. by replacing old information with the new ones.

**Limitation on the Evaluation and Analysis**

Secondly, how to evaluate and analyze common grounding largely remains an open problem. One main difficulty of evaluation is that systems can *imitate* common grounding without actual understandings, either through simple acknowledgements and paraphrasing (Weizenbaum, 1966) or based on more elaborate human-like responses (Adiwardana et al., 2020; Roller et al., 2021). Such deceptive behaviors could be misleading or even harmful to promote research on reliable common grounding. Another difficulty is the existence of *dataset biases*. In the process of dataset creation, there could be various sources of suprious, unintended biases (Goyal et al., 2017; Gururangan et al., 2018; Geva et al., 2019) which can be exploited by the models to succeed without employing the genuine, intended abilities (Geirhos et al., 2020). To overcome these limitations, we need more objective, quantitative and faithful evaluation metrics to measure and propel our real progress on advanced common grounding.

In terms of analysis, there are various factors that make the analysis of common grounding difficult or problematic in realistic settings. For instance, the application of Clark and Schaefer's theory may not be straightforward when contributions are implicit, indirect, unstructured, partial, etc. To illustrate this, we use another example from Sacks et al. (1974), following the explanations of Lascarides and Asher (2009).

(2) a. Mark (to Karen and Sharon): Karen 'n I're having a fight,
b. Mark (to Karen and Sharon): after she went out with Keith and not me.
c. Karen (to Mark and Sharon): Well Mark, you never asked me out.

In this dialogue, Mark and Karen agree on the fact that the *cause* of the fight was Karen going out with Keith: however, this is only presented through an *implicature* in Mark's utterances (from the "fight" occurring after Karen "went out with Keith"). Furthermore, Karen accepts this without direct acknowledgement, but instead through the initiation of a relevant contribution (i.e. the *explanation* of why she went out with Keith). This suggests that both the presentation and acceptance phase of the contribution can be implicit and difficult to identify in actual conversations. Hence, we need more reliable methods to interpret and analyze the intermediate process of common grounding.

**Limitation on the Model Capability**

Finally, we raise the limitation of the model capability for sophisticated common grounding. Generally speaking, there are two mainstream approaches in building contemporary dialogue systems (Gao et al., 2019). The first approach is the so-called *traditional* approach, where the dialogue system is carefully designed as a pipeline of specialized modules. In this approach, each module would be responsible for a certain subprocess of dialogue, such as parsing user utterances (Yao et al., 2013), tracking dialogue states (Henderson et al., 2014b), planning next utterances (Peng et al., 2017) and realizing their surface forms (Wen et al., 2015). While this approach has the advantage of being more interpretable, controllable and robust to expectable errors, the reliance on manual design (e.g. on the modularization) becomes a bottleneck in achieving scalability, generality and flexibility required for human-level common grounding.

The second approach is the so-called *end-to-end* approach, where the dialogue system is developed in a fully data-driven manner with minimal prior constraints, typically

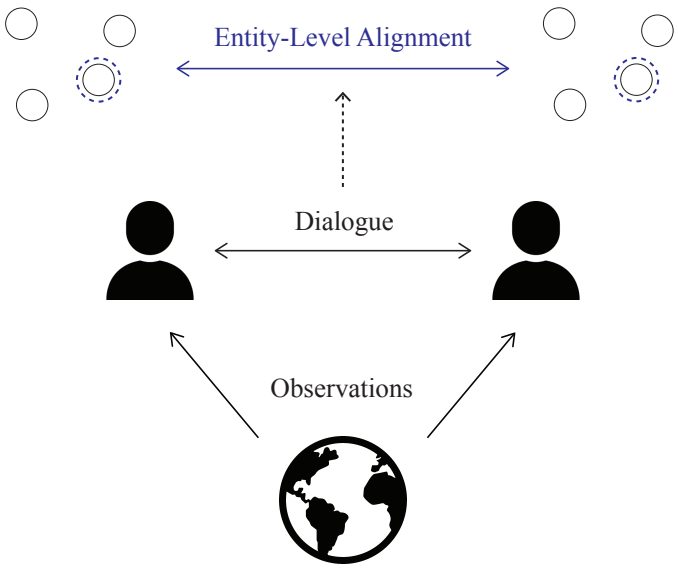



Figure 1.1: An illustration of our task formulation as *entity-level alignment*.

using the neural networks (Vinyals and Le, 2015; Bordes et al., 2017). As long as sufficient/appropriate data are available, this approach has the advantage of being more scalable, flexible and applicable to general domains: in fact, even to the most challenging *open domains* as required for the chatbots (Adiwardana et al., 2020; Roller et al., 2021). However, the major drawback is the lack of interpretability, controllability and robustness to recover in the face of even trivial mishaps arising from miscommunication (Brennan, 1998; Benotti and Blackburn, 2021).

Therefore, incorporating human design (first approach) and learning directly from data (second approach) have complementary strengths/weaknesses, and neither is sufficient by itself. While there are recent approaches that aim the ideal middle ground (Williams et al., 2017; Andreas et al., 2020), realizing human-level common grounding (e.g. in terms of both flexibility and robustness) still remains an open challenge.

## 1.3 Contributions of the Thesis

To address all of these problems, we propose a novel research platform to study advanced common grounding in natural language dialogue systems. To make the scope of this thesis precise, we focus on the aspect of *entity-level alignment* as the critical first step in general common grounding. As visually depicted in Figure 1.1, this step can be formalized as the following dialogue process:

1. First, we assume that each speaker has a private *observation* of the environment. This can be respresented as a personal experience, knowledge, perception, etc, typically involving multiple entities in the environment.

2. Secondly, the speakers converse in natural language to recognize the same entity in the environent. The type of entity to be recognized can vary depending on the objective, e.g. physical objects, locations, or temporal events.

3. Finally, we consider common grounding to be *successful* if and only if the speakers could recognize the same entity. In other words, we consider common grounding to be an accurate *alignment* of the private observations at the *entity-level*.

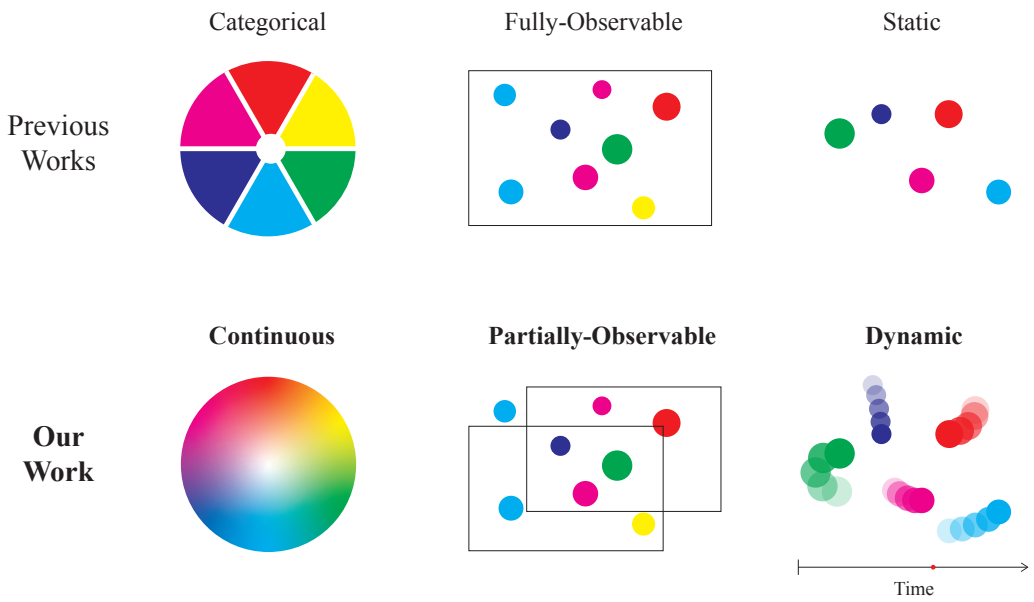



Figure 1.2: An illustration of the task settings in previous works (top row) and our work (bottom row). We introduce *continuity* to require more semantic coordination, *partial observability* to require the resolution of potential misunderstandings, and *dynamics* to incorporate the aspect of maintaining/updating common ground.

For instance, the example dialogue (1) can be considered as the alignment of a *locational* entity and example (2) as the alignment of an *event* (or possibly its *cause*). While this formalization is simplified and ignores the aspect of entire common grounding, entity-level alignment is a critical and indispensable step in any type of common grounding, e.g. developing common ground of an entire scenery (Das et al., 2017a; Haber et al., 2019; Alamri et al., 2019).

By restricting our scope to this preliminary setting, we address each of the three major limitations in the following ways.

### Contribution on the Task Settings

We propose a novel task setting under *continuous*, *partially-observable* and (optionally) *dynamic* context to require advanced skills of common grounding: see Figure 1.2 for a visual illustration in comparison to the previous settings. To be specific, we represent the dialogue context based on *continuous* real values, which reflects the information (e.g. perception) in the physical world more faithfully and requires more advanced coordination, e.g. based on nuanced expressions (Paradis, 2008) and pragmatic reasoning (Monroe et al., 2017). In addition, we introduce *partial-observability* where the agents have different perspectives of the environment, which is more realistic and requires to take into account the possibility of various misunderstandings (Keysar et al., 2000). Finally (as an optional setting), we incorporate diverse *dynamics* of the environments to require advanced skills of maintaining common ground: concretely, we incorporate dynamic scenes based on animations to require spatio-temporal reasoning (Girdhar and Ramanan, 2020) as well as information updates to require adaptation to the changing environments (Moon et al., 2020).

It is worth noting that each of these settings has been explored at least partially or independently in the previous literature: for instance, De Vries et al. (2017) proposed a reference game under *continuous* visual context, and He et al. (2017) worked on a *partially-observable* setting. However, our important contributions are the *abstraction* of

the task settings in the three universal dimensions and their *combinations* to require more advanced common grounding. Owing to this level of abstraction, we can empirically investigate the consequences of each task setting in a scientific manner (i.e. through controlled experiments and hypothesis testing) and expect our findings to be general and fundamental for designing experimental setups in future dialogue research.

**Contribution on the Evaluation and Analysis**

To evaluate the ability of creating common ground, we formulate a novel *collaborative reference task*, where the goal of the agents is to select (i.e. coordinate attention on) the same entity through natural language dialogue. We consider this as an important step in general common grounding, since the mutual recognition of discourse entities (i.e. *entity-level alignment*) is a fundamental building block in successful communication. Under this task formulation, we can evaluate the ability of accurate common grounding based on the *task success rate*, which is both quantitative and objective. We also focus on a *synthetic* environment to control, balance and diversify the dialogue contexts: this way, we can minimize undesirable biases and enable more faithful evaluation (Johnson et al., 2017; Girdhar and Ramanan, 2020). We show an example dialogue of our collaborative reference task under *continuous* and *partially-observable* (but not *dynamic*) context in Figure 1.3 (top), which will be formally introduced in Chapter 3.[2]

To analyze the process of common grounding, we propose a simple, reliable and useful framework of linguistic annotations. Specifically, we first identify the *referring expressions* in the dialogue and the corresponding *referent entities*, which can be in another modality (such as vision). This allows us to interpret the intermediate process of common grounding, such as how misunderstandings are introduced and resolved. Secondly, we conduct additional annotations to capture the detailed strategies of common grounding, such as how referring expressions are predicated or omitted in ellipses. To this end, we focus on *spatial expressions* which are prevalent in visually grounded dialogues (including our own). We show an illustration of our analytical framework in Figure 1.3 (middle), which will be discussed in detail in Chapter 4 and 5. Note that our analyses span different modalities (language and vision), which allow for investigating the related and important problem of *symbol grounding* (c.f. Section 2.3).

Under dynamic context, the collaborative reference task can be temporarily generalized to track and select the same entity at *multiple timesteps* in the same environment: we refer to this as the *sequential* collaborative reference task. Under this task formulation, we can further evaluate the ability of *maintaining* accurate common ground based on the length of successful timesteps. We show an example dialogue of our sequential collaborative reference task in Figure 1.3 (bottom), which will be formally introduced in Chapter 6.

Based on the proposed task formulations and analytical frameworks, we enable reliable evaluation and detailed analysis of the fundamental aspects of common grounding.

**Contribution on the Model Capability**

Following the above ideas, we developed a collection of large-scale resources to train, evaluate and analyze various dialogue systems in terms of advanced common grounding. To be specific:

- First, we collected 6,760 dialogues based on our minimal collaborative reference task under *continuous* and *partially-observable* context (Figure 1.3, top): we refer

[2]Note that we also kept the *task complexity* minimal for the ease of analysis.

to this as OneCommon Corpus. Based on this dataset, we evaluate our baseline model's capability of *recognizing* the created common ground in Chapter 3.

- Secondly, we curated 5,191 successful dialogues from OneCommon Corpus and conducted the annotation of reference resolution (Figure 1.3, middle left). We describe how this resource can be leveraged to interpret and improve the common grounding strategies of our baseline model in Chapter 4.

- Thirdly, we randomly sampled 600 dialogues (already annotated with reference resolution) and further conducted the annotation of spatial expressions (Figure 1.3, middle right). Based on this annotation, we assess how well our improved baseline can recognize the fine-grained linguistic structures of the visually-grounded dialogues in Chapter 5.

- Finally, we collected additional 5,617 dialogues based on our sequential collaborative reference task under *continuous*, *partially-observable* and *dynamic* context (Figure 1.3, bottom): we refer to this as Dynamic-OneCommon Corpus. Based on this dataset, we assess our baseline model's capability of creating, retaining and updating common ground in dynamic environments in Chapter 6.

In this thesis, we mainly focus on the simple and widely used *end-to-end* dialogue systems (Vinyals and Le, 2015; Bordes et al., 2017; Lewis et al., 2017) as our baselines. However, our platform can be leveraged to experiment with various other approaches, as we discuss in Section 2.2. While we leave further model assessments and improvements as future work, we propose a fundamental testbed for evaluating, analyzing and improving dialogue systems in terms of advanced common grounding.

To summarize our contributions, we proposed novel task settings to study advanced common grounding as entity-level alignment. Based on our careful task designs, we formulated the (sequential) collaborative reference tasks to evaluate the ability of creating and maintaining accurate common ground. We also proposed useful analytical frameworks to interpret and analyze the intermediate process of common grounding. Finally, we developed large-scale resources for conducting various empirical studies, including the evaluation, analyses and improvements of data-driven dialogue systems.

Overall, we expect our proposed platform to be fundamental for promoting research on advanced common grounding in natural language dialogue systems.

## 1.4 Thesis Outline

The outline of this thesis is as follows:

- In Chapter 2, we give an overview of the existing literature related to common grounding. This includes its theoretical foundations, computational research, and relationship with symbol grounding approached within various fields. Whenever appropriate, we clarify/explicate the novelty, motivations and contributions of this thesis to place them in these broader contexts.

- In Chapter 3, we introduce a novel collaborative reference task under *continuous* and *partially-observable context*. Based on this task formulation, we collected 6,760 dialogues through crowdsourcing, which we refer to as OneCommon Corpus. Through our dataset analysis and experiment, we verified the advanced common grounding strategies required in this setting as well as the open room left for improvement. This chapter is based on our published work (Udagawa and Aizawa, 2019).

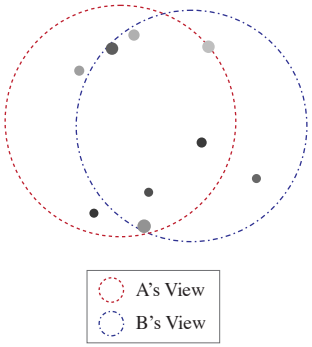
- In Chapter 4, we propose a method of interpreting common grounding based on *reference resolution*. Based on our novel framework, we annotated 5,191 dialogues from OneCommon Corpus via a combination of expert annotators and crowd workers. Our dataset analysis and experiment demonstrate the advantages of our annotation for interpreting the intermediate process of common grounding. This chapter is based on our published work (Udagawa and Aizawa, 2020).

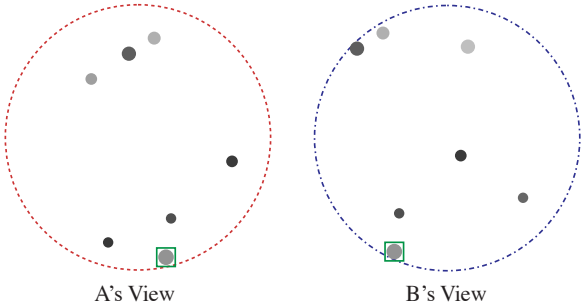
- In Chapter 5, we conduct further analyses by leveraging the existing annotation (reference resolution) to annotate *spatial expressions*. We capture fine-grained linguistic structures of 600 dialogues in OneCommon Corpus, including predicate-argument structure, modification and ellipsis. Based on our improved baseline, we run a comprehensive assessment of the model's capability for recognizing such structures. This chapter is based on our published work (Udagawa et al., 2020).

- In Chapter 6, we propose a novel task setting under *dynamic* context to study the ability of mainining common ground. Based on our *sequential* collaborative reference task, we crowdsourced 5,617 human dialogues, which we refer to as Dynamic-OneCommon Corpus. Through our dataset analysis and experiment, we demonstrate even more sophisticated strategies required in this setting in comparison to the *static* counterpart of OneCommon Corpus. This chapter is based on our work to be published (Udagawa and Aizawa, 2021).

- In Chapter 7, we discuss the promising directions worth exploring in future research. Specifically, we propose further ideas on the *task design methodologies* to require fully advanced common grounding, *model improvements* to achieve more flexible and robust common grounding, and *real-world applications* that naturally follow from the studies in this thesis.

## Chapter 3: Task Formulation




A: I see three in a line going up and to the right. The middle one is the largest and darkest

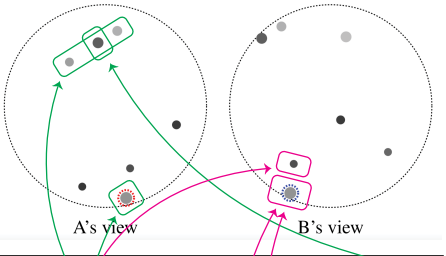
B: I don't see that. I have one large, medium gray dot that's under a small, darker gray dot

A: Is the larger dot slightly to the left

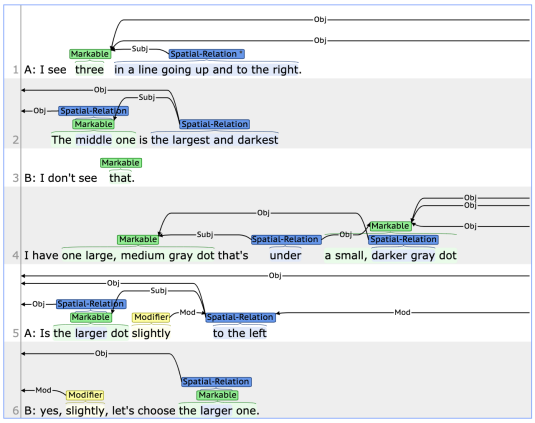
B: yes, slightly, let's choose the larger one

A selects: B selects:

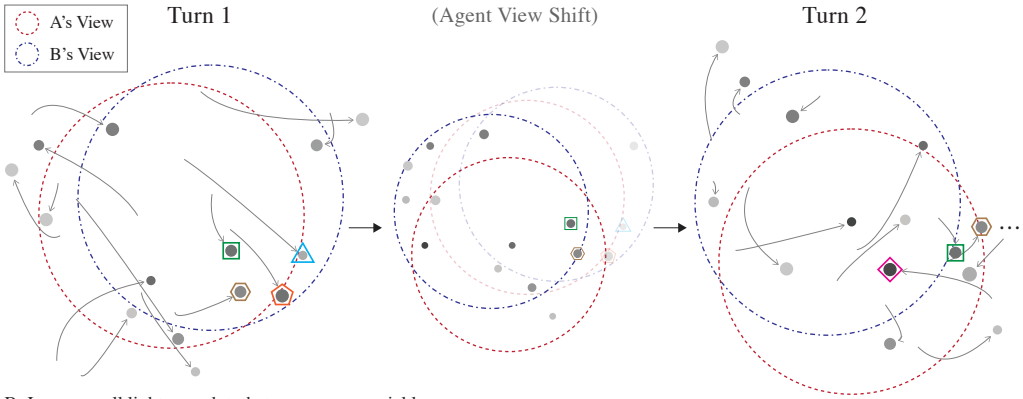
## Chapter 4 & 5: Interpretation and Analysis




A: I see three in a line going up and to the right. The middle one is the largest and darkest
B: I don't see that. I have one large, medium gray dot that's under a small, darker gray dot
A: Is the larger dot slightly to the left
B: yes, slightly, let's choose the larger one
A: SELECT red
B: SELECT blue




## Chapter 6: Task Generalization




B: I see a small light grey dot, that moves very quickly. It ends to the right of three larger, darker dots. ( )

A: i think that one moves off my view. do you have two medium sized dark dots on top of one another at the start, with the upper one being a bit smaller and off to the left? they both move down and to the right ( )

B: Yes, I think so. But the one that started off on the left moves very slowly? And another lighter grey one ends up almost in between them? ( )

A: yes can you pick the slower moving one?

A selects: B selects:

A: do you still see the same one?

B: Nope, it barely squeaked out of view. But a larger black ball comes towards the left into view. Do you see it? It probably crosses underneath our old one. ( )

A: yes

A selects: B selects:

Figure 1.3: An overview of our task formulations and analytical frameworks.

# Chapter 2

# Literature Review

Due to the foundational aspect of common grounding, there is a rich and diverse literature related to this topic. In this chapter, we first give an overview of the existing literature on common grounding (Section 2.1). Then, we review the contemporary approaches to dialogue system engineering (Section 2.2). Finally, we discuss the important links to the related (yet distinct) problem of symbol grounding (Section 2.3).

## 2.1 Common Grounding

In this section, we first give an overview of the theoretical foundations of common grounding, including the literature in philosophy, mathematical logic and psycholinguistics (Section 2.1.1). Then, we review the computational models of common grounding in natural language conversations, including both formal and empirical approaches (Section 2.1.2).

### 2.1.1 Theoretical Foundations

The notion of common ground (or *common knowledge*) was originally conceived by David Lewis (Lewis, 1969) to explain the nature of human *conventions*. Conventions are the regularities in our behavior which we expect each of us to understand and follow: without them, we cannot account for our solutions to various *coordination problems* (Schelling, 1960), from setting up daily meetings to communicating in natural language. In essence, Lewis argues that such expectations can only be justified based on a certain level of common ground established within the community.

In later studies, common ground came to be recognized as ubiquitous and essential in various linguistic phenomena/competence, including presupposition (Stalnaker, 1978), definite reference (Clark and Marshall, 1981), communication (Clark, 1996) as well as first language acquisition (Clark and Grossman, 2001).

But what exactly is common ground? Can it be defined formally, and do we know how exactly they are built up? The absolute answers may not be established yet, but there are three major approaches which account for these questions.

**Epistemic Logic** The first approach is studied in the field of epistemic modal logic (Fagin et al., 2003). To give a brief introduction, we consider $A$ to be the set of agents and $p$ to be a proposition. Then, using the modal operator $K_a$, we can represent the epsitemic state that "an agent $a \in A$ knows $p$" as $K_a\, p$. Similarly, we can represent the state that "all agents in $A$ know $p$" based on the operator $E_A\, p$ (which is equivalent to $\bigwedge_{a \in A} K_a\, p$). Finally, if we use the abbreviation $E_A^n\, p$ for $E_A\, E_A^{n-1}\, p$ (where $E_A^0\, p = p$), we can represent the state "$p$ is *common knowledge* among $A$" based on the operator $C_A\, p$ (which is equivalent to $\bigwedge_{i=1}^{\infty} E_A^i\, p$).

The process of common grounding can be taken into account by using additional operators which represent the *actions* or *events* that cause the change in agents' epistemic states: such logical framework is known as dynamic epistemic logic (Van Ditmarsch et al., 2007). If we assume the agents to be perfectly logical reasoners, this framework allows for computing the exact set of common ground after each information update.

However, the formulations and assumptions in this approach may not be cognitively plausible. For instance, a literal interpretation of common knowledge imposes an infinite list of epistemic conditions to be checked (Clark and Marshall, 1981), and humans may not be perfect logical reasoners either. This drawback makes the theory difficult to be applied in realistic settings, such as natural language dialogues.

**Shared Basis** The second approach defines common ground based on the notion of *shared basis* (Lewis, 1969; Clark and Marshall, 1981). To be specific, this approach considers a proposition $p$ to be common ground among the agents i.f.f. there exists some state of affairs $b$ such that:

1. Every agent has information of the basis $b$.
2. $b$ indicates to every agent that every agent has information of $b$.
3. $b$ indicates to every agent that $p$ holds.

For instance, in a situation where agents are *co-present* and *co-observing* a house on fire, these state of affairs become the shared basis for the common ground "there exists a house on fire". This approach is more cognitively plausible because the infinite, recursive assessment of epistemic states is not required (as in the previous approach). Furthermore, the shared basis need not be represented symbolically and hence accounts for *multi-modal* (non-linguistic) common grounding through gestures (Lascarides and Stone, 2009), eye gaze (Nakano et al., 2003) and other mediums that are commonly used in face-to-face interactions.

However, current approaches so far define common ground independently for each proposition $p$. Is it possible to define common ground (and common grounding) in its entirety, including their interrelationships and non-propositional knowledge? The last approach offers an interesting perspective which takes this into account.

**Situation Models** The last approach employs the concept of *situation models*. Situation models are the mental representations of the state of affairs described in text or conversation (Zwaan and Radvansky, 1998). For instance, there is empirical evidence in cognitive psychology that humans are sensitive to spatial layouts when comprehending stories (Glenberg et al., 1987) and constructs spatially coherent representations as they read along (Ferguson and Hegarty, 1994). These studies suggest that comprehension of language is not merely an analysis of syntactic/semantic structures, but should rather be considered as construction of situation models coherent with the linguistic descriptions (e.g. in terms of spatial, temporal and causal dimensions).

Following this idea, Pickering and Garrod (2004) consider common ground(ing) as the *alignment* of situation models among the interlocutors. Based on this view, the process of common grounding does not require explicit reasoning of the epistemic states, as required in the previous approaches. Instead, they argue that such alignment proceeds mostly implicitly at various linguistic levels, including phonetic, lexical, syntactic and semantic representations. For instance, lexical entrainment naturally leads to aligned conceptualization, which is also in line with the theory of *conceptual pacts* (Brennan and Clark, 1996).

This approach has the advantage of representing the entire common ground as the *aligned situation models*, which capture the holistic view of common ground (including their internal structures and non-linguistic knowledge). However, the lack of strict definitions and strong empirical evidence (e.g. on how situation models are actually aligned) potentially remain as its major challenges.

It is worth noting that our formalization of common grounding (Section 1.3) is closest to the last definition by Pickering and Garrod (2004). However, we circumvent the difficulty of dealing with entire situation models by restricting our scope to *entity-level alignment*, which keeps the task simple, well-defined and quantifiable.

### 2.1.2 Computational Approaches

The previous section focused on the conceptual (*epistemic logic*) or psychological (*shared basis* and *situation models*) foundations of common grounding. Unfortunately, they are not directly applicable for our practical goal of developing *computational* models of common grounding through *natural language* interaction.

#### Formal Approaches

To model common grounding in actual human conversations, Traum (1994) conducted a pioneering work based on a refinement of the contribution theory (Clark and Schaefer, 1989). To be specific, he defined a set of *grounding acts* (such as *initiate*, *continue*, *repair*, *request repair* and *acknowledge*) by taking into account the speech acts (Austin, 1962; Searle, 1969) relevant to common grounding. Based on these grounding acts, a finite state model is used to compute the transition of information in dialogue, until it reaches the final accepting state (and considered to be in common ground).

In subsequent works, the idea of this approach has been integrated with Discourse Representation Theory (Kamp, 1981; Kamp and Reyle, 1993) to represent the semantic (e.g. propositional) contents of common ground (Poesio and Traum, 1997). In Segmented Discourse Representation Theory (Asher et al., 2003), *rhetorical relations* are also taken into account to capture the implicit process of common grounding, e.g. through the initiation of the next relevant contribution. Such formal representations provide elaborate computational models of not only common grounding but natural language dialogue in general (Ginzburg, 2012).

However, formal approaches come with their own disadvantages. For instance, identifying the precise semantic representation of a dialogue is a challenging task requiring time and high expertise. They are also unable to capture various complexities in realistic dialogues, such as ambiguity and uncertainty in common grounding. Finally, they are not directly applicable to *situated* dialogues involving rich non-linguistic contexts, such as vision and embodied environments.

#### Empirical Approaches

To overcome these limitations, recent works mostly focus on empirical methods and develop *data-driven* models of common grounding. The primary step in this approach is to collect dialogue corpora: such attempts have originated in the HCRC Map Task Corpus (Anderson et al., 1991) and continued since then (Potts, 2012; Tokunaga et al., 2012; Zarrieß et al., 2016). Recent works often utilize crowdsourcing to collect diverse data at scale: notable works include the GuessWhat?! dataset (De Vries et al., 2017), MutualFriends dataset (He et al., 2017), PhotoBook dataset (Haber et al., 2019) as well as our own (Udagawa and Aizawa, 2019, 2021). As an illustration, we show an excerpt

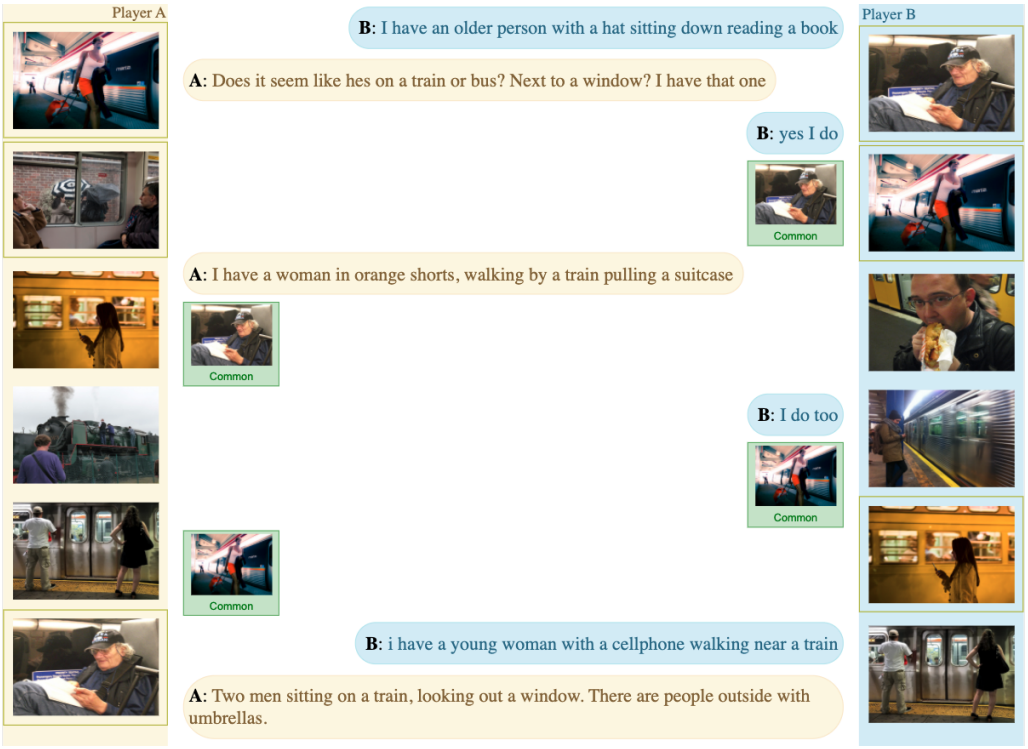



Figure 2.1: An excerpt from the PhotoBook dataset (Haber et al., 2019). In this task, two human players are provided different combinations of images and try to find out whether the (bordered) images are in common or not.

of the common grounding task from the PhotoBook dataset in Figure 2.1, where the human players try to identify the common images through dialogue.[1]

The second step is employ machine learning (Murphy, 2012) to train the models based on the collected corpora, typically in an *end-to-end* or fully data-driven manner. By leveraging the actual human dialogues, we can expect the models to learn fundamental strategies of common grounding (e.g. clarifications and elaborations) without expensive, potentially unreliable annotations. We can also expect the models to capture the ambiguity and uncertainty in common grounding owing to their high flexibility. Finally, this approach can naturally be applied in situated conversations, e.g. by treating multi-modal information in a unified manner based on neural networks (Goodfellow et al., 2016). We will explain the specific model architectures in Section 2.2.3.

Note that most of the existing common grounding tasks focus on closed domain, goal-oriented dialogues, including our collaborative reference tasks (c.f. Figure 1.3): this is for the sake of quantitative and objective evaluation based on the *task success rates*, as we discussed in Chapter 1. However, some recent works also attempt to model this in open domain, non-goal-oriented dialogues: for instance, Cho and May (2020) developed a model of (single-turn) common grounding through creative improvisation of the next relevant contribution. While its evaluation remains an open challenge, this new approach sheds light on interesting potential applications, such as the development of more engaging, entertaining and coherent chat-bots.

[1]Note that the players are indeed relying on *entity-level alignment* to succeed in common grounding.

## 2.2 Dialogue Systems

In this section, we review the broader literature of dialogue system research in general. Unfortunately, most of the existing works do not specifically focus on common grounding: for instance, they do not require advanced common grounding skills nor enable direct evaluation/analysis of such abilities. However, understanding the broader literature is fundamental for common grounding research, since they provide crucial insights on the related problems and engineering solutions. Besides, it is important to place the role of common grounding in the general settings of dialogue, which will be one of the main focuses of this section.

### 2.2.1 Tasks and Datasets

Dialogue research is conducted under various motivations and practical goals: consequently, numerous tasks and datasets have been proposed in the existing literature. To make our review as inclusive as possible, we categorize them into the following 4 types: namely *(traditional) task-oriented dialogues*, *multi-modal dialogues*, *non-cooperative dialogues*, and *non-goal-oriented dialogues*.

**Task-Oriented Dialogues** Traditionally, the central motivation of dialogue system research has been a practical one: to develop industrial-level dialogue systems that can execute a variety of useful tasks. The expected tasks include booking a trip (El Asri et al., 2017) or recommending movies (Chen et al., 2019a) by interactively soliciting the user's preferences. Due to its practicality, many existing benchmarks and shared tasks have been focusing on such task-oriented dialogues (Henderson et al., 2014a; Budzianowski et al., 2018; Lee et al., 2019). Recent works propose valuable extensions as well, e.g. based on scalable task ontologies (Rastogi et al., 2020) and efficient domain adaptations (Shalyminov et al., 2020)

However, the underlying settings in these tasks remain mostly unchanged and do not require advanced common grounding. To be specific, they only need to deal with *categorical* databases or knowledge graphs, which do not require semantic coordination from the ground up. For instance, continuous concepts like the price range would be discretized in advance (e.g. to "cheap", "moderate" and "expensive"), so that they can avoid sophisticated common grounding and keep the tasks manageable.

If we were to develop truly conversational and practical systems, we should start focusing on systems that can deal with more advanced settings, e.g. that can negotiate on the exact preferences on a *continuous* scale. Without such ability, dialogue systems cannot adequately respond to the user's precise requests, e.g. in terms of the price range, location, atmosphere, and so on.

**Multi-Modal Dialogues** Recently, increasing attention has been paid to multi-modal dialogues, especially those grounded in the visual modality (Das et al., 2017a; De Vries et al., 2017; Alamri et al., 2019). These works lie at the intersection of computer vision (CV) and NLP, involving intricate cognitive reasoning (Zellers et al., 2019) and spatio-temporal grounding (c.f. Section 2.3.2). Such extension to the visual modality is also useful in practice. For instance, Moon et al. (2020) proposed a dialogue task where the system can display and manipulate a shared screen with the user: this allows for better interaction and improved user experience, especially in shopping domains (e.g. for fashion and furniture) where item apperances play a key role.

However, it remains unclear to what extent existing tasks require precise dialogue understanding and spatio-temporal grounding. First, there are minimal linguistic analyses conducted on these datasets, and the complexities of the dialogue structures remain

unrevealed. Secondly, there could be various sources of undesirable biases in realistic visual contexts (Goyal et al., 2017; Cirik et al., 2018), and the tasks may not necessarily require genuine intricate abilities. For instance, existing visual dialogue tasks may be largely solvable without taking into account *dialogue history* (Agarwal et al., 2020), contrary to the expectation when it was first proposed (Das et al., 2017a).

Therefore, we argue that more attention must be paid to the requisite skills of visually-grounded dialogues and their reliable evaluation. Common grounding and symbol grounding are the two central competence we focus on and address in this thesis.

**Non-Cooperative Dialogues** All of the previous tasks so far (including the common grounding tasks) have assumed that dialogue is *cooperative*, i.e. the goals of the interlocutors are fully aligned. However, this may not necessarily be the case in general. Typical departures include the cases of negotiation (Lewis et al., 2017; He et al., 2018; Li et al., 2020), where the interlocutors' goals may be different or even conflicting (e.g. they must compete for a fixed amount of resource). Such situations may also arise when one needs to persuade the other to profit a third party (Wang et al., 2019) or in the case of malicious scams where one is intending to deceive the other (Li et al., 2020). On such occasions, the interlocutors need to exert complex reasoning skills and negotiation strategies to achieve their desired outcomes.

However, this does *not* indicate that common grounding can be non-cooperative in such settings. On the contrary, we argue that fully cooperative common grounding plays a fundamental role in maintaining any type of conversation. Even in the case of negotiation, the interlocutors must cooperate and share (certain levels of) the same understandings to disagree, compromise, or take any extreme measures (Lewis, 1969; Sacks, 1987). This suggests that we must take into account the collaborative aspect of common grounding, even if our goal was to study dialogues in non-cooperative settings.

**Non-Goal-Oriented Dialogues** Finally, we raise the category of *non-goal-oriented* dialogues, typically referred to as chit-chats. The main objective of this research is to develop *engaging* and *coherent* chatbots that can handle a wide variety of topics, ranging from sports, politics to technology (Ram et al., 2018). One key factor that affects engagement is *empathy*, the cognitive ability of understanding and acting on behalf of the user's emotional states. Microsoft XiaoIce, one of the world's most popular chatbots today, explicitly incorporates such empathetic mechanisms (Zhou et al., 2020). To assess and improve the consistency of chatbots, Zhang et al. (2018) proposed a dialogue setting where the interlocutors are provided artificial profiles to act in accordance to. In subsequent works, the profiles are extended to fictitious character settings and grounded in more open-ended text adventure games (Urbanek et al., 2019).

In recent years, there has been an impressive breakthrough on the model side based on gigantic-scale neural networks (Adiwardana et al., 2020; Roller et al., 2021). While they exhibit almost human-level performance under specific setups, they still do not possess the true competence of natural language conversation. To demonstrate this, Benotti and Blackburn (2021) raised an example dialogue with BlenderBot, the state-of-the-art chatbot from Facebook (Smith et al., 2020; Roller et al., 2021):

(1)
a. Human: Do you know Blender?
b. BlenderBot: Yes, I've used it a little bit. I am not very good at it, though.
c. Human: But I don't mean the 3D software.
d. BlenderBot: Oh, I see. Yeah, I know a little bit about it. What do you like to do with it?
e. Human: I mean Blender, the chatbot. Have you used it?
f. BlenderBot: Oh, yes, I have used it. I like to use it to make 3D models.

In this conversation, the human speaker clarifies that "Blender" refers to the chatbot and not the 3D software (c). Impressively, BlenderBot acknowledges its misunderstanding and makes a perfectly coherent recovery (d). However, in the last utterance (f), it turns out that BlenderBot had not actually resolved its misunderstanding and once again misinterprets "Blender" as the 3D software.

This example demonstrates the lack of true common grounding capacity of modern chatbots: in this case, the ability of accurate *entity-level alignment*. To realize truly engaging and coherent chatbots, we expect that more robust and reliable common grounding will be the key in non-goal-oriented dialogues as well.

### 2.2.2 Evaluation Metrics

Evaluation plays a crucial role in the development of dialogue systems. The specific measures can vary depending on the task domain: for instance, the *task success rate* may be the primary metric in task-oriented dialogues, *each player's payoff* may be so in non-cooperative dialogues and *user satisfaction* in non-goal-oriented dialogues. Such evaluation can be conducted either manually (e.g. based on questionnaires) or automatically (e.g. based on regressions from the relevant factors, Walker et al. 1997).

There are also task agnostic approaches to dialogue evaluation. For instance, *human-likeness* is usually a desirable factor in all types of dialogues. While this is relatively straightforward to measure manually based on human judgements, automatic evaluation of such quality remains an open challenge. For instance, reference-based metrics such as BLEU (Papineni et al., 2002) may not correlate with human evaluation (Liu et al., 2016), and regression-based metrics such as ADEM (Lowe et al., 2017) can easily be fooled by simple adversarial attacks (Sai et al., 2019).

For a comprehensive review of existing evaluation methodologies, we refer to a recent survey by Deriu et al. (2020). However, throughout this thesis, we argue that the evaluation of dialogue systems should focus more on their *abilities* (Hernndez-Orallo, 2017) rather than the ad-hoc performance metrics (such as benchmark scores). This means that we should answer the following research questions through the design of dialogue tasks and evaluation metrics:

- What are the specific abilities required to succeed in this task setting?
- Does the evaluation metric truly reflect such abilities?

In the field of psychometrics, these two questions roughly correspond to the concepts of *validity* and *reliability*, respectively (Cook and Beckman, 2006). Unfortunately, current evaluation of dialogue systems often neglects the first question: for instance, it is not made clear what linguistic capabilities and strategies are required in a certain dialogue setting. Many works also fail to answer the second question: for instance, the existence of dataset biases makes the evaluation unreliable and unfaithful.

From this aspect, we hope this thesis contributes important insights on the requisite skills and evaluation of *common grounding*, which should be taken into account in all types of dialogues (as discussed in Section 2.2.1).

### 2.2.3 Model Architectures

Finally, we give an overview of the mainstream architectures in dialogue system engineering. As briefly discussed in Chapter 1, there are two major approaches in developing modern dialogue systems. In Figure 2.2, we show an illustration of the two approaches from a recent survey paper (Gao et al., 2019).

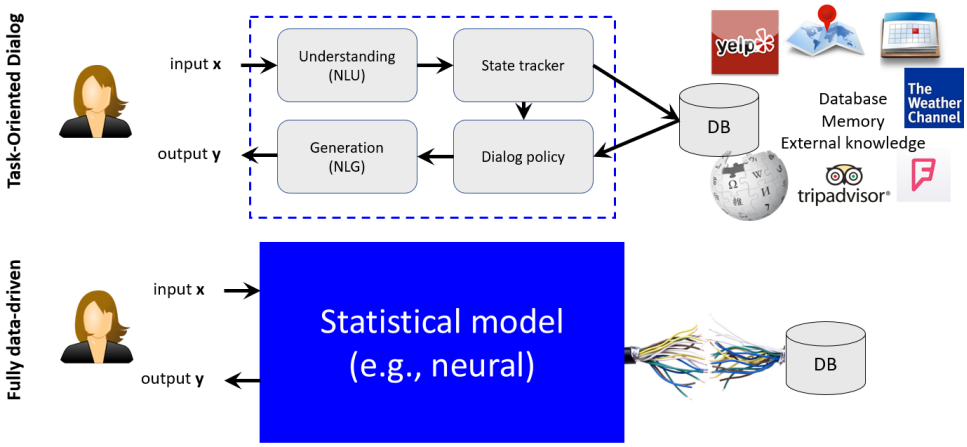



Figure 2.2: An illustration of the two mainstream architectures of data-driven dialogue systems from Gao et al. (2019).

**Traditional Dialogue Systems** The first approach is based on a pipeline/modular architecture, which has traditionally been used for task-oriented dialogue systems (Figure 2.2, top). In this framework, the *natural language understanding (NLU) module* is first applied to the user utterance to extract relevant information, typically in a structured format. This can be as simple as a pair of user intent and named entities (Gupta et al., 2006) or more complex graph-based representations which capture anaphora, relations, and utterance ambiguity (Kollar et al., 2018). Next, the *dialogue state tracker* integrates the extracted information with the previous dialogue state to update the current state. Dialogue states are often represented based on the *frame* structure (i.e. slot-value pairs), possibly one for each domain under control (El Asri et al., 2017). In a recent approach, these are extended to executable data-flow graphs (Andreas et al., 2020) which are much more general and expressive. Based on the updated dialogue state, the system can optionally query the database for necessary information (Dhingra et al., 2017) and the *dialogue policy module* determines the content of the system response (Takanobu et al., 2019). Finally, the *natural language generation (NLG) module* either retrieves an appropriate response template (Yan et al., 2016) or generates the surface form conditioned on the determined content (Wen et al., 2015).

**End-to-End Dialogue Systems** In the second approach, dialogue systems are typically trained in a fully data-driven manner based on *end-to-end* neural networks. Under this setting, dialogue modeling is often treated as a type of *(conditional) language modeling* problem (Vinyals and Le, 2015): conditioned on the input sequence of dialogue history, compute the probability of the next dialogue responses.

Formally, let $D = \{d^1, d^2, ..., d^{|D|}\}$ be the dialogue corpus containing the dialogue sequence $d^i = \{x_1^i, x_2^i, ..., x_{|d^i|}^i\}$. We generally consider each element of the dialogue sequence $x_t^i$ to be a *token*: however, it can also be in the unit of the whole *utterance* (Serban et al., 2016). Given $D$ as the dataset, the learning objective is to find the model parameters $\theta$ that minimize the sum of negative log probabilities (or equivalently, maximize the product of total probabilities)

$$L(\theta) = \sum_{i=1}^{|D|} \sum_{t=1}^{|d^i|} -\log p(x_t^i | x_{<t}^i, \theta) \tag{2.1}$$

where $x_{<t}^i = \{x_1^i, x_2^i, ..., x_{t-1}^i\}$ (which represent dialogue history).

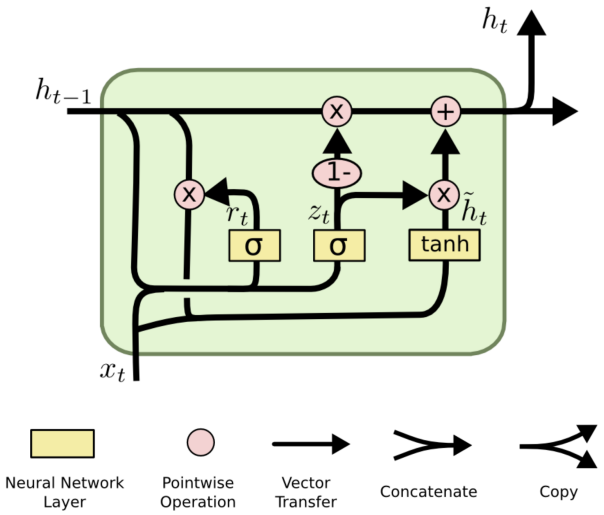



Figure 2.3: An illustration of the Gated Recurrent Unit (GRU) architecture, from the online post by Olah (2015).

This can be modeled by a variety of deep learning architectures in the *encoder-decoder* framework. Specifically, we first encode the dialogue history $x^i_{<t}$ with an *encoder*, followed by a *decoder* to predict $x^i_t$. There are multiple architectures that can handle this, including LSTM (Hochreiter and Schmidhuber, 1997), GRU (Cho et al., 2014) and Transformers (Vaswani et al., 2017). Arguably, the simplest architecture is the GRU (short for "Gated Recurrent Units"), which is illustrated in Figure 2.3. This is a type of recurrent neural network, whose hidden state $h_t$ at position $t$ is dependent on the input $x_t$ and previous state $h_{t-1}$ and computed in the following way:[2]

$$
\begin{aligned}
z_t &= \sigma\left(W_z \cdot (h_{t-1} \odot x_t)\right) \\
r_t &= \sigma\left(W_r \cdot (h_{t-1} \odot x_t)\right) \\
\tilde{h}_t &= \tanh\left(W \cdot ((r_t * h_{t-1}) \odot x_t)\right) \\
h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t
\end{aligned}
$$

Here, $W_z$, $W_r$ and $W$ represent learnable matrices, $\cdot$ represents the matrix (dot) product, $\odot$ represents the vector concatenation, and $*$ the elementwise (Hadamard) product. Crucially, $h_t$ is a linear interpolation of the previous state $h_{t-1}$ and the updated state $\tilde{h}_t$ (based on $x_t$) to address the *vanishing gradient problem* (Bengio et al., 1994).

As a last remark, we can easily incorporate (non-linguistic) dialogue context into the problem formulation. To be specific, let $D = \{(d^1, c^1), (d^2, c^2), ..., (d^{|D|}, c^{|D|})\}$ represent the dialogue corpus containing the additional dialogue context $c^i$. Then, the learning objective would be to minimize the following loss function:

$$
L(\theta) = \sum_{i=1}^{|D|} \sum_{t=1}^{|d^i|} -\log p(x^i_t | x^i_{<t}, c^i, \theta) \tag{2.2}
$$

This can be handled by a straightforward extension of the encoder, e.g. using Convolutional Neural Networks (CNN) to encode visual contexts (Krizhevsky et al., 2012; Xu et al., 2015). Due to its flexibility, scalability and generality, this approach has been applied in a wide variety of tasks, including traditional task-oriented dialogues (Bordes et al., 2017; Wen et al., 2017; Ham et al., 2020), multi-modal dialogues (Das et al.,

[2]Note that we abbreviate the superscript $i$ (indicating the ID of the dialogue sequence) for simplicity.

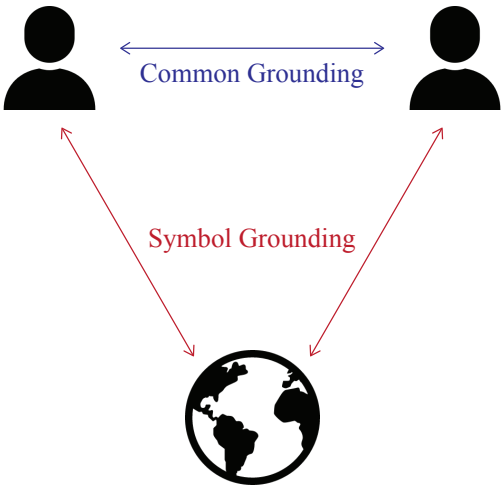



Figure 2.4: The triadic relations of common grounding and symbol grounding.

2017c; Kottur et al., 2018; Lu et al., 2019a), non-cooperative dialogues (Lewis et al., 2017; Yarats and Lewis, 2018; Li et al., 2020) and non-goal-oriented dialogues (Vinyals and Le, 2015; Adiwardana et al., 2020; Roller et al., 2021).

In both approaches, common ground is represented as the *dialogue state* of the systems: namely *structured* representations (such as frames and data-flow graphs) in the first approach and *learned* representations (e.g. the hidden state $h_t$) in the second approach. Due to this difference, the first approach tends to perform more reliably in schematic traditional dialogue tasks (Takanobu et al., 2020), while the second approach is preferred in other domains. Since our proposed tasks are non-schematic and require dealing with non-linguistic (visual) contexts, we generally focus on the second approach (end-to-end dialogue models) as our baselines in this thesis.

## 2.3 Symbol Grounding

*Symbol grounding* is another primary competence of natural language, which refers to the process of connecting symbolic language to the physical world (Harnad, 1990). In this section, we first discuss the important relations between common grounding and symbol grounding with respect to *natural language acquisition* (Tomasello, 2009). Secondly, we focus on computational approaches to *spatio-temporal grounding* and review the emerging literature in the fields of CV and NLP.

### 2.3.1 Language Acquistion

The meaning of language cannot be defined by the language itself, as it would lead to an endless circular definition: instead, it must be somehow connected to the physical world in the right way (Harnad, 1990). At the same time, we cannot use language in an arbitrary way if we were to communicate with each other: the usage of language must become a *convention* through the process of common grounding (Lewis, 1969). These two requirements form an intricate triadic relationship in terms of constructing and acquiring natural language, as depicted in Figure 2.4.

This relationship has been noticed and investigated in a variety of literature. In developmental psychology, the establishment of *joint attention* (mutual recognition of an external entity) has been regarded as a crucial basis for first language acquisition (Clark and Grossman, 2001; Tomasello, 2009): this is a representative example where common

grounding meets symbol grounding. In the fields of linguistics and AI, there has been an intense debate over the schools of nativism, empiricism and culturism (Chomsky, 1957; Elman et al., 1996; Steels, 1997). For instance, the naming of color in human language may be affected by the innate structure of our perception, the physical world and environments we experience, or the requirements as a social tool for communication (Steels and Belpaeme, 2005). The converging view seems to suggest that all of these perspectives matter, complexly involving innate human structures, symbol grounding and common grounding (Larsson, 2018).

On the other hand, the NLP community has been predominantly focusing on the study of static corpora. Such corpus-driven approach originates in the seminal work of Penn Treebank (Marcus et al., 1993) and often validated on the account of the *distributional hypothesis*, known for the famous quote from Firth (1957) – "you shall know a word by the company it keeps". This approach has been successful to a remarkable extent: state-of-the-art models achieve near human-level performance on a range of natural language understanding benchmarks (Wang et al., 2018a; Devlin et al., 2019), even in the challenging few-shot settings (Brown et al., 2020). However, upon deeper inspection, their linguistic abilities often turn out to be brittle and shallow, as we've seen in the case with BlenderBot (c.f. Section 2.2.1). Consequently, increasing works caution that language acquisition cannot be realized only from its forms (corpora) and requires interaction with the real world and other humans (Bisk et al., 2020; Bender and Koller, 2020; Benotti and Blackburn, 2021).

Building upon these backgrounds/observations, we make an important contribution to study the triadic relationship of common grounding and symbol grounding in a minimal yet advanced setting. To be specific, we focus on the primary notion of *joint attention* and incorporate realistic factors of human dialogues in the physical environments: namely *continuity*, *partial-observability* and *dynamics*.

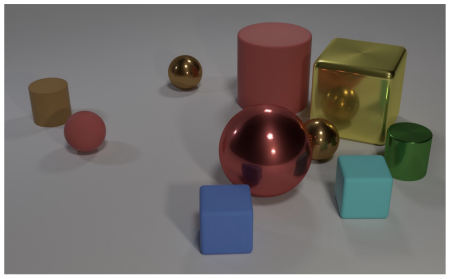
### 2.3.2 Spatio-Temporal Grounding

While symbol grounding encompasses all conceivable connections between language and the physical world, one principal connection is made on the dimensions of space and time. This type of connection, which we refer to as *spatio-temporal grounding*, has been an important subject of study at the intersection of CV and NLP.

**Visual Grounding Tasks** Vision-language grounding has been a longstanding goal in the fields of CV and NLP. This is fundamental for developing systems that can not only *speak* but also *perceive* and *act* in the physical world, just as we normally do. Such technology is also crucial for making information accessible to the visually-impaired, e.g. by automatically rendering visual contents to natural language descriptions.

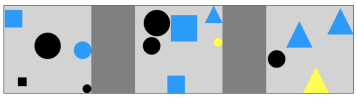
One representative task of visual grounding is the *visual question answering (VQA)* task. This requires the models to answer correctly to the questions (information requests) related to the visual context, including images (Antol et al., 2015; Hudson and Manning, 2019) as well as videos (Lei et al., 2018, 2020). In the *visual captioning (VC)* task, the goal is to provide the descriptions of salient objects, events, background, etc in a given scene through natural language (Xu et al., 2015; Hossain et al., 2019). Finally, in *visual dialogues*, the task is extended to an interactive setting, typically involving both the *reactive* aspects of VQA and *proactive* aspects of VC (Das et al., 2017a; De Vries et al., 2017; Alamri et al., 2019).

However, realistic visual contexts are often unbalanced and lack diversity in terms of spatio-temporal relations. For instance, in natural photographs, a dog would most likely be *under* a tree rather than *on top of* a tree. In natural videos, a book would typically *fall off* the shelf but not *float up* to the shelf. When such biases exist, the spatio-temporal



**Q:** Are there an equal number of large things and metal spheres?
**Q:** What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
**Q:** How many objects are either small cylinders or metal things?

*There are two towers with the same height but their base is not the same in color.*



*There is a box with 2 triangles of same color nearly touching each other.*

Figure 2.5: Example questions from the CLEVR dataset (Johnson et al., 2017) (left) and the NLVR dataset (Suhr et al., 2017) (right).

attributes and relations may be identifiable based on superficial clues only, e.g. by simply recognizing the object categories (*dog*, *tree*, *book*, *shelf*, etc) (Goyal et al., 2017; Cirik et al., 2018).

One solution to circumvent this issue is to use synthetic visual contexts. As an illustration, we show example instances from CLEVR (Johnson et al., 2017) and NLVR datasets (Suhr et al., 2017) in Figure 2.5. In these datasets, the entity attributes and relations are restricted to preliminary concepts but randomly sampled in a fully controlled manner. This way, their *combinatorial* variety is maximized while reducing exploitable biases, requiring genuine understanding and compositional generalization (Fodor and Pylyshyn, 1988; Lake and Baroni, 2018). Furthermore, the controllability of the context allows for various in-depth analyses of complex models, e.g. error analyses and *causal* analyses through direct intervention (Pearl, 2009; Pearl and Mackenzie, 2018).

While these two approches have complementary strengths and weaknesses, we generally follow the second line of research (based on synthetic contexts) to maximize the diversity of spatio-temporal expressions and enable faithful evaluation and diagnosis of spatio-temporal grounding.

**Spatio-Temporal Language** Finally, spatio-temporal language has a long history of research as well. In computational linguistics, annotation frameworks have been proposed to capture the events, time and temporal relations expressed in natural language, e.g. in news articles (Pustejovsky et al., 2003; Ning et al., 2018). The annotation of spatial language has also been attempted to capture the information related to locations, directions, topological relations and motions (Pustejovsky et al., 2011a; Dan et al., 2020). The recognition of such spatio-temporal information plays a fundamental role in natural language comprehension (Zwaan and Radvansky, 1998) as well as pratical applications such as information extraction and text mining (Banko et al., 2007; Aggarwal and Zhai, 2012).

In the CV community, one of the most popular representations of spatio-temporal information is based on the *scene graphs* (Johnson et al., 2015; Krishna et al., 2017b). In this approach, spatio-temporal attributes and relations are captured in the *predicate-argument structure*, as shown in Figure 2.6. This allows for a natural mapping from spatio-temporal language to *regions* (e.g. object boundaries) in the visual scenes.

However, existing works mostly focus on *monologues* rather than dialogues, ignoring the important intersection between symbol grounding and common grounding. In dialogues, speakers often have incomplete/asymmetric information, requiring collaborative coordination on the semantics, references, *frames* of references, etc: for instance, the referent of the speaker may not be visible/groundable for the listener. To address such

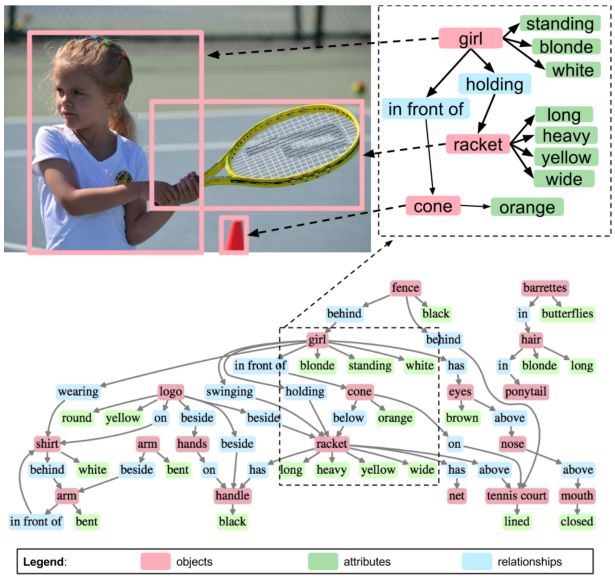



Figure 2.6: An example of the scene graph representation from Johnson et al. (2015).

issues, interlocutors often rely on *modifications* to express precise information (Paradis, 2008) and *ellipses* to make conversation efficient.

These factors can make the application of existing approaches (e.g. annotation frameworks and scene graphs) nontrivial, yet mostly ignored in the prior literature (e.g. Elliott and Keller 2013). In this thesis, we explicitly take into account such complexities and propose a simple yet challenging testbed of spatio-temporal language in interactive conversations.

# Chapter 3

# Task Formulation under Continuous and Partially-Observable Context

Common grounding is the process of creating, repairing and updating mutual understandings, which is a critical aspect of sophisticated human communication. However, existing dialogue systems still have limited capability of creating common ground, and we also lack task formulations which introduce natural difficulty of common grounding while enabling easy evaluation and analysis of complex models. In this chapter, we propose a minimal dialogue task which requires advanced skills of common grounding under *continuous* and *partially-observable* context. Based on this task formulation, we developed OneCommon Corpus: a large-scale dataset of 6,760 dialogues which fulfills essential requirements of natural language corpora. Through our dataset analysis, we uncover the difficulty of common grounding and other relevant phenomena that need to be considered. Finally, we evaluate and analyze baseline neural models on a simple subtask that requires accurate recognition of the created common ground. Our results show that the baseline models perform decently but leave room for further improvement.

## 3.1 Introduction

One major goal of NLP is to develop systems with human-level competence of dialogue. In the field of psychology, the ability to construct and maintain common ground has been pointed out to be essential for natural language communication (Clark, 1996) as well as acquisition (Tomasello, 2009). Furthermore, in human-computer interaction (HCI), it is important that humans and computers have certain ways of creating *mutual understandings* to collaborate reliably (Brennan, 1998). Although natural language is not the only option, it is one of the most effective, effortless and versatile solutions to this problem.

### Problems of Existing Research

However, existing study of common grounding in dialogue system research is limited in three major ways. First, existing dialogue tasks are limited in terms of common grounding due to the restricted types of information that need to be dealt with. Specifically, they are focused on either *categorical* or *fully-observable* context, which makes common grounding a relatively trivial task:

- Under *categorical* context (Bordes et al., 2017; He et al., 2017; Lewis et al., 2017), information can be expressed by symbolic natural language with minimal ambiguity. For example, there could be little ambiguity in describing categorical attributes like discrete colors ("red", "blue" and "yellow"). However, under *continuous* context, natural language usage becomes more nuanced and pragmatic (such as "darker gray" and "almost black") for precise semantic coordination.

- Under *fully-observable* context (Zarrieß et al., 2016; De Vries et al., 2017), it is usually given that every information about the context is shared among the agents. This makes common grounding easier because information of the context is already present in (or at least easily added to) their common ground. In contrast, under *partially-observable* context, agents typically need to create common ground from minimal shared information, and there could be various misunderstandings or partial understandings that need to be resolved.

Another problem is the difficulty of evaluation and analysis. As the models acquire more flexibility, automatic evaluation becomes problematic (Liu et al., 2016; Novikova et al., 2017) and interpretation of model behavior becomes more challenging. Since advanced common grounding requires high flexibility, it is expected that we will need reliable evaluation metrics and analysis methods in the process of comparing and improving different methods.

Finally, there are limitations of model capabilities. Although traditional dialogue systems rely on rule-based engineering and predetermined slot-filling (Traum, 1994; Young et al., 2013; Williams et al., 2016), these models lack flexibility in terms of representing dialogue states and generating natural utterances. Since common ground can be very complex involving high ambiguity, uncertainty, partial understandings and misunderstandings, we need systems that can better capture such complexities and resolve them through flexible dialogues.

### Our Contributions

In this chapter, we make a first step towards addressing these problems in the following ways. First, we formulate a novel dialogue task which requires advanced skills of common grounding under *continuous* and *partially-observable* context. Our task is based on a more general *collaborative reference task*, where the goal of the agents is to coordinate attention on the same entity in a given context. This setting enables clear evaluation based on the task success rate and various error analyses of arbitrary models, since the contexts are simple and completely controllable (Section 3.2).

Second of all, to enable the development of recent end-to-end dialogue systems with high flexibility (Bordes et al., 2017; Lewis et al., 2017), we collected a large-scale dataset of 6,760 human-human dialogues with over 32K utterances through crowdsourcing on Amazon Mechanical Turk. During the dataset collection, we defined and managed to fulfill three essential requirements of natural language corpora: namely interpretability, linguistic/strategic variety and reliability (Section 3.3).

Next, we conduct comparative analyses with the previous dataset to illustrate how continuous and partially-observable context introduces difficulty in terms of common grounding. In addition, we conduct further analyses of the dataset to investigate the common grounding strategies at different levels, including nonlinguistic bias towards *joint saliency* (Section 3.4).

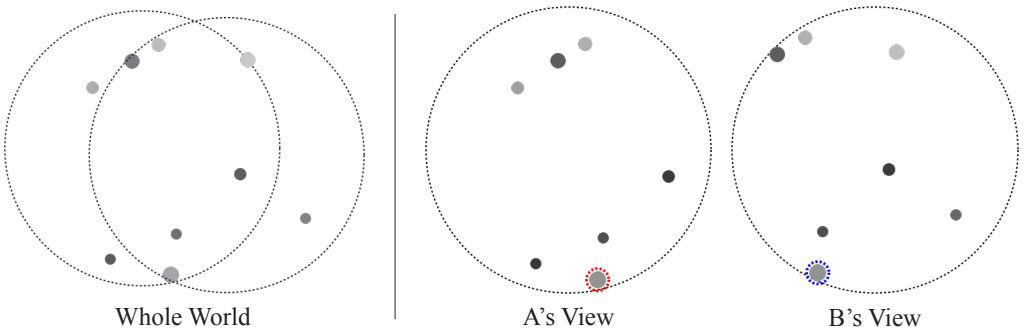
Finally, we evaluate and analyze simple neural network models on our dataset based on an important subtask of collaborative reference. Due to the complexity of common grounding, there remains huge room for further improvement (Section 3.5).

Overall, the contributions of this chapter are as follows:

- We proposed a simple and general idea of incorporating continuous and partially-observable context to the dialogue tasks, which makes common grounding difficult in a natural way.
- Following this idea, we formulated a minimal collaborative reference task which enables clear evaluation and analysis of complex models.

- We collected a large-scale dataset of 6,760 dialogues, which fulfills essential requirements of natural language corpora.
- Our analysis of the dataset verified the difficulty of common grounding and revealed various phenomena that need to be considered.
- We evaluated simple baseline models on an important subtask of collaborative reference and demonstrated huge room left for improvement.

## 3.2 Task Formulation




A: I see three in a line going up and to the right. The middle one is the largest and darkest
B: I don't see that. I have one large, medium gray dot that's under a small, darker gray dot
A: Is the larger dot slightly to the left
B: yes, slightly, let's choose the larger one.
A: SELECT red
B: SELECT blue

Figure 3.1: An example dialogue of our collaborative reference task. Two agents have partial (overlapping) observations of a 2-D plane with 7 entities in each view. Their goal is to find and select the same entity through natural language communication.

A *collaborative reference task* is a multi-agent cooperative game with entities $E = \{e_1, e_2, ..., e_m\}$ and agents $A = \{a_1, a_2, ..., a_n\}$. Each agent $a_i \in A$ has an observation of $E$, and they can freely exchange information with other agents in natural language. At the end of the game, each agent selects one of the observable entities, and the game is considered *successful* if and only if all the agents selected the same entity. This is a general framework for evaluating accurate *mutual recognition* of a common entity, which is often a critical step in general common grounding.

Note that in contrast to the typical reference tasks (Kazemzadeh et al., 2014; De Vries et al., 2017), agent roles are *symmetric* and they can agree on any of the common entities (as long as it's the same). A dataset closest to our setup is the MutualFriends dataset (He et al., 2017), which is based on the task of finding a mutual friend from private lists of friends. Although this can be considered as a collaborative reference task under *partially-observable* context (due to the privacy of knowledge), they only include *categorical* information and the difficulty of common grounding is limited.

Based on this task formulation, we propose a minimal task setting under *continuous* and *partially-observable* context. To be specific, we consider two agents and multiple entities located on a 2-D plane. Each entity has 3 simple attributes (2-D location, size and color) represented as *continuous* real values. Furthermore, the agents are located slightly differently on the plane, and they can only observe the entities within a fixed radius: this way, our setting is made *partially-observable* as well.

We show an example dialogue of our task in Figure 3.1. Although human players successfully coordinated their selections with a short number of utterances, we can verify advanced common grounding strategies such as pragmatic expressions ("three in a line going up"), clarifications based on hypothesis testing ("Is the larger dot slightly to the left") and nuanced acknowledgements ("yes, slightly").

For the sake of simplicity, the number of entities observable by each agent is fixed at 7. This ultimately reduces our task to a simple *classification* problem, which can be evaluated based on simple metrics (such as accuracy).

## 3.3 Dataset Collection

We basically followed the dataset collection procedure of the MutualFriends dataset (He et al., 2017). We used Amazon Mechanical Turk to pair up 2 crowd workers and gave 20 seconds to read an example dialogue, followed by a maximum of 6 minutes session to complete the task. Our chat interface is shown in Figure 3.2.

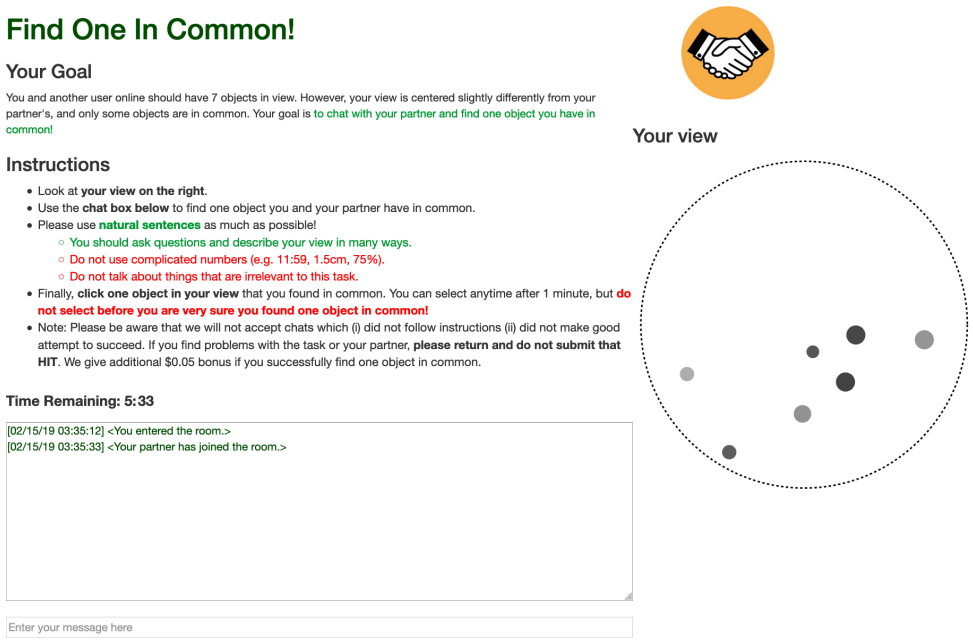



Figure 3.2: A screenshot of our chat interface. Workers are given a maximum of 6 minutes to identify one of the common entities through dialogue.

In this procedure, we were concerned with three essential requirements of natural language corpora: namely *interpretability*, *linguistic/strategic variety* and *reliability*. In the following, we discuss the importance of each requirement as well as how we managed to fulfill them:

**Interpretability** We define the interpretability of the dataset to be the ease of interpreting its language and strategy, which is critical for conducting annotations and further analyses. However, *lack of discipline* and *complexity of the vocabulary* can make corpora difficult for interpretation.

In free-formed dialogues, lack of discipline can cause unnecessary difficulty in terms of interpretation. For instance, *cross-talk* (conversation which does not progress sequentially) can occur frequently (He et al., 2017) and complicate important structures of dialogues, such as discourse segments, adjacency pairs and *contributions* in common grounding (Clark and Schaefer, 1989). Hence, we minimized them by forcing workers

to take turns. Also, *chit-chat* could occur occasionally, which adds undesirable noise for analyzing common grounding. Therefore, we explicitly instructed the workers to avoid talking about things that are irrelevant to this task.

Keeping the vocabulary simple is also important for interpretability, especially for people unfamiliar with the domain. For instance, the MutualFriends dataset includes up to 7 attributes with approximately 3K complex named entities and technical terms. In contrast, we kept the entity attributes minimal with only 3 simple and intuitive properties (location, size and color). As a result, this greatly reduced the complexity of the vocabulary as we describe in Section 3.4.

**Linguistic and Strategic Variety** *Linguistic and strategic variety* of the dataset is fundamental for developing dialogue systems with broad coverage. To satisfy this requirement, we sampled all entity attributes randomly from fixed uniform distributions, with the only restriction that the entities cannot be too close to each other. As the previous work with the similar idea confirmed (Suhr et al., 2017), we found rich varieties of linguistic phenomena, including cardinalities ("*three* gray dots"), existentials ("*There is* another small dark .."), universals ("*all* of the other dots are larger"), coordinations and negations ("further to the right *and not* as far down").

During the dataset collection, we assigned 6,759 unique contexts to 6,760 dialogues we collected. We collected 2 dialogues based on the exact same context and confirmed that they solved the task in different ways. This suggests that there could be various effective solutions and the agents must adapt flexibly to their partners' strategies.

Finally, we incorporated variation in terms of the *degree* of partial observability. Specifically, each agent has 7 entities in the view, but only 4, 5 or 6 of them would be in common. This introduces a further variation of common grounding strategies, as we discuss in detail in Section 3.4.

**Reliability** Finally, we regard the *reliability* of the dataset to be crucial, especially when crowdsourcing data with potentially careless, low-motivated workers. In fact, in our preliminary experiment, we found many cases where workers did not follow the instruction carefully or solve the task effectively (especially on difficult cases).

As a solution, we manually reviewed all works and rejected ones which clearly did not follow the instruction. Our instruction is kept brief and explicit so that it is easier to follow, and we also gave manual feedback about general solutions to improve their work. To discourage premature guessing, we prohibited workers from selecting within the first minute and instructed them to make it *very sure* they found the same entity before selection. We also incentivized task success with $0.05 bonus for all successful dialogues, in addition to the base reward of $0.30.

As a result, we found significant improvement in terms of task success rate, which is an important evidence of the reliability of our dataset.

Based on the above procedure, we collected 6,839 dialogues, and after the reviewing process, we accepted 6,760 dialogues in total. We refer to this dataset as **OneCommon Corpus**. Overall, we received positive feedback from the crowd workers and they seemed to enjoy playing it, which is important from the aspect of gamification.

## 3.4 Dataset Analysis

In this section, we first study the difficulty of common grounding in comparison to the previous settings. Secondly, we conduct further analyses to investigate other relevant phenomena that need to be considered.

### 3.4.1 Difficulty of Common Grounding

Our hypothesis is that *continuous* and *partially-observable* context makes common grounding difficult compared to *categorical* or *fully-observable* context. However, it is relatively obvious that *fully-observable* setting would make collaborative reference trivial: for instance, if our task was fully-observable, one can easily succeed by always uttering "*select the leftmost dot*". Therefore, we mainly focus on assessing how *continuous* context adds difficulty in terms of common grounding.

As a comparison, we use the MutualFriends dataset which is based on a similar collaborative reference task under partially-observable but *categorical* context (He et al., 2017). However, several differences make a direct comparison difficult: for instance, we only gave one chance for the entity selection, while the MutualFriends dataset allowed multiple chances in the given amount of time. Therefore, we focus on the following factors which are less affected by such differences.

#### Utterance Lengths

| | MutualFriends | OneCommon | | |
|---|---|---|---|---|
| | | #Shared=4 | #Shared=5 | #Shared=6 |
| Total dialogues | 10,661 | 2,189 | 2,279 | 2,292 |
| Average tokens per utterance | 5.38 | 12.87 | 12.37 | 11.86 |
| Average utterances per dialogue | 8.97* | 4.97 | 4.77 | 4.56 |
| Success rate (%) | 0.85* | 0.66 | 0.77 | 0.87 |
| Unique tokens | 13,478 | | 3,761 | |
| Occupancy of top 10% frequent tokens (%) | 91.6% | | 97.0% | |

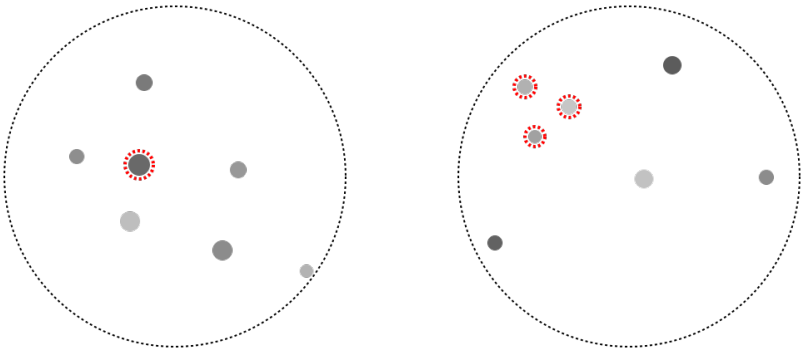
Table 3.1: Basic statistics of our dataset and the MutualFriends dataset (He et al., 2017). To count tokens and vocabulary size, we preprocessed the text with the same NLTK word tokenizer (Loper and Bird, 2002) and converted each token to its lowercased form. (* denotes where direct comparison is not suitable due to the task difference.)

First, we compare the *average utterance length* because this indicates the syntactic/semantic complexity of utterances required for common grounding. As shown in Table 3.1, utterances in our dataset are at least twice as long compared to the MutualFriends dataset: this indicates that more complex utterances are required under continuous context to make the descriptions precise. We also found that the utterance lengths slightly increase when the number of shared entities is smaller: this suggests that (a greater degree of) partial-observability also adds complexity at the utterance level.

#### Pragmatic Expressions

In our dataset, we found many *pragmatic expressions* whose meaning depends on the context and should not be taken literally. A typical example is the usage of the word "black" to indicate the darkest dot in the context, even if its color is not completely black. Another common expression "triangle" is also pragmatic, since in the literal sense there could be numerous triangles in one's view, and the speaker actually indicates a group of three dots which is the closest to a *prototypical* type of triangles, such as an equilateral triangle. We show some illustrating examples in Figure 3.3.

As the previous work pointed out, such pragmatic expressions are characteristic under continuous context (Monroe et al., 2017) and add complex ambiguity/uncertainty that need to be resolved through common grounding.



large **black** dot a close **triangle**

Figure 3.3: Examples of typical pragmatic expressions in our dataset (marked by bold).

### Nuanced Expressions

Finally, the frequent usage of *nuanced* expressions is an important characteristic of our dataset. Since the context is continuous and partially-observable, we hypothesize that speakers need to rely on such expressions to express subtle differences in terms of degree, ambiguity, uncertainty, and so forth.

To estimate the frequency of nuanced expressions, we follow Paradis (2008) and define 2 main types (and 5 subtypes) of degree modifiers: *scalar modifiers* used for concepts in a range of scale (*diminishers*, *moderators*, *boosters*) and *totality modifiers* used for concepts with definite boundaries (*approximators*, *maximizers*). See Figure 3.4 for the proposed taxonomy and examples of such degree modifiers.

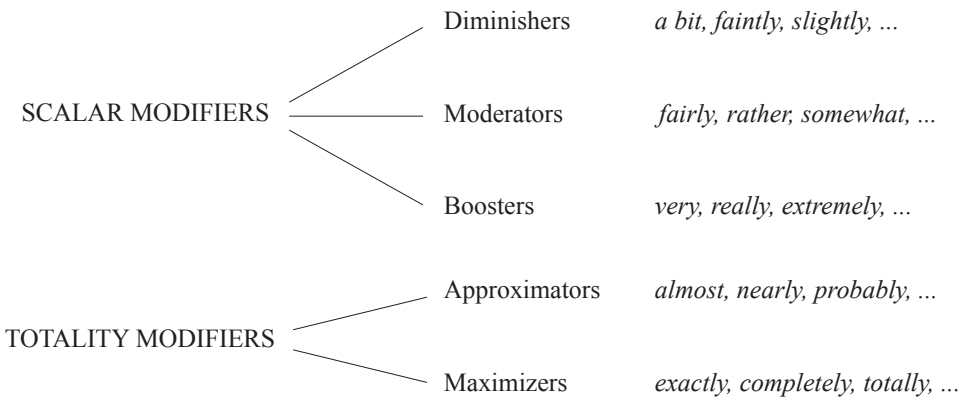



Figure 3.4: A taxonomoy of the degree modifiers based on Paradis (2008).

Based on this taxonomy, we manually created a keyword-based dictionary of the degree modifiers. Note that we excluded words which are likely to be used with different meanings (such as "like", "about" and "around") and do not consider nuances expressed morphologically (such as *-ish* as in "small*ish*"), although they are also common in our dataset. Using this dictionary, we estimated the frequency of the modifiers in the MutualFriends dataset and OneCommon Corpus. As shown in Table 3.2, we can verify that our dataset includes significantly more degree modifiers of various types, which are used effectively to cope with ambiguity and uncertaintiy.

To summarize our analyses, utterances are much longer in our dataset, which indicates the complexity of common grounding at the utterance level. In addition, our dataset

| Degree Modifiers | MutualFriends | OneCommon | # Keywords | Usage in OneCommon |
|---|---|---|---|---|
| Diminishers | 0.01 | 9.19 | 10 | **slightly** to the right |
| Moderators | 0.09 | 1.28 | 6 | **fairly** close together |
| Boosters | 0.17 | 9.83 | 27 | **very** light dot |
| Approximators | 0.86 | 10.23 | 34 | **almost** in the middle |
| Maximizers | 0.47 | 4.31 | 37 | **exactly** horizontal |

Table 3.2: Average occurrences of degree modifiers per 100 utterances (estimated based on keywords).

involves more ambiguity and uncertainty represented by the frequent usage of pragmatic and nuanced expressions. These observations verify that common grounding becomes more difficult and complicated under continuous and partially-observable context.

On the other hand, we found that human workers could solve the task reasonably well with little evidence of confusion. Therefore, we conclude that our task setting is fundamental for adding natural difficulty in terms of common grounding.

### 3.4.2 Other Relevant Phenomena

Next, we conduct further analyses of the dataset and investigate various phenomena related to common grounding that need to be considered.

#### Basic Statistics

From Table 3.1, we also found that dialogues get longer in terms of the *(average) number of utterances* when a fewer number of entities is shared. This shows that under a greater degree of partial-observability, it is more likely that the presented information is not *groundable*, and players need more interaction (try-and-error) to create common ground. We can also see that the success rate drops naturally, so in general common grounding becomes more challenging when less information is shared.

In terms of lexical variety, our dataset contains 3,761 unique tokens in total, in contrast to 13,478 in the MutualFriends dataset. In addition, we found that a large portion of our dataset consists of common words: to be specific, the top 10% of the most frequent tokens occupy 97.0% of the whole tokenized corpus, in contrast to 91.6% in the MutualFriends dataset. These suggest that the vocabulary of our dataset is extremely simple, which is an important evidence of interpretability discussed in Section 3.3. This may also be helpful for training dialogue systems, since rare words are less problematic.

#### Nonlinguistic Phenomena

Language is a *coordination device* we use to coordinate our joint actions (Lewis, 1969), but we also use *joint saliency* to coordinate actions at the nonlinguistic level (Schelling, 1960). In our dataset, we found that human players have a tendency to focus attention on *perceptually* salient entities more often.

To demonstrate this, we plot the final selection probabilities of the entity's color and size in Figure 3.5. We can clearly see that the selections are biased, and entities with extreme properties (around the edge) are more likely to be selected. We also found that darker entities are more likely to be selected (62.7%) compared to lighter entities (37.3%), and larger entities (54.3%) slightly more likely than smaller entities (45.7%).

There could be other types of joint saliency as well (such as geometric relations between entities), but the point is that such bias exists and needs consideration: for

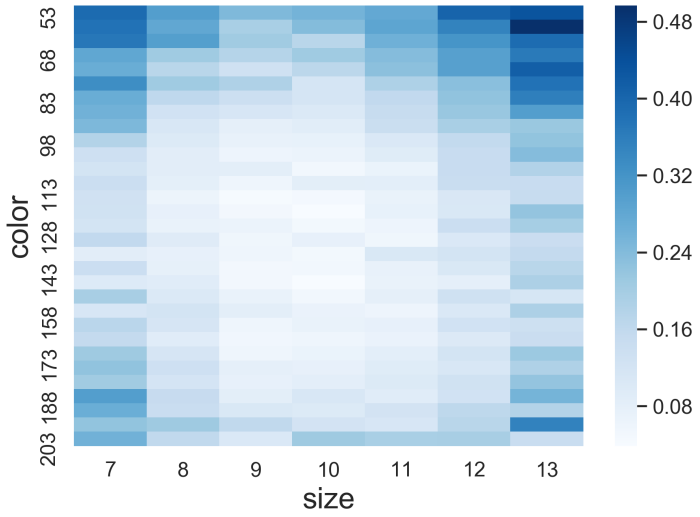



Figure 3.5: Final selection probabilities based on color and size. The total range of size is 7 and color is split into 30 equal-sized bins in grayscale (smaller is darker).

example, just by taking advantage of such bias, we can predict human selections significantly better than random (Section 3.5). However, due to the *partially-observability*, joint saliency is not sufficient and communication is critical to solving our task.

### Utterance Level Phenomena

Understanding speaker's intention at the *utterance level* is critical in dialogue (Grice, 1957): especially the idea of speech act (Austin, 1962; Searle, 1969) has been applied widely in dialogue system research to improve natural language understanding and generation (Yao et al., 2013; Hakkani-Tür et al., 2016).

During data collection, we allowed free-formed conversation with minimal restrictions, as long as they are relevant to the accomplishment of the task. As a result, we found a wide variety of speech acts in the course of common grounding. We show illustrative examples of the collected utterances in Table 3.3. Utterances are grouped by the *communicative functions* defined in Bunt et al. (2017), including *information transfer functions* (*information provision/request*) and *action discussion functions* (*commissives/directives*). With additional annotations, our dataset can be easily extended to study such pragmatic structures as well (Stolcke et al., 2000).

### Discourse Level Phenomena

Finally, we found many coreferences and anaphoric expressions in our dataset. *Coreference resolution* is the task of finding the mentions (referring expressions) in the dialogue with the same referents (Ng, 2010). In our dataset, we found two characteristics that complicate this task. First, due to the continuous and partially-observable context, mentions are usually ambiguous and the referents could be missing from one's view. Hence, players must keep track of various possibilities and disambiguate them through interaction. Secondly, players often use the *grouping* strategy (such as "three in a line", "a cluster of 4 dots") where the mention can refer to a *set* of multiple entities. This is a natural and effective strategy but adds complexity in terms of coreference resolution.

On the other hand, *anaphora resolution* is the task of idenfitying the relation between a mention (*antecedent*) and a succeeding mention (*postcedent*) that depends on the pre-

| Function | Type | Example Utterances |
|---|---|---|
| Info. Provision | Inform (Init.) | I have very dark small dot in the center |
| | Inform (Cont.) | It also has a small light grey one further down from the group |
| | Agreement | Yes I have one like that. / same here. |
| | Agreement (Strong) | Exactly! / perfect. mine too. |
| | Agreement (Partial) | not sure its the one / more of a line.<br>Yes, but the small is medium dark, not completely black |
| | Disagreement | I don't have that one. / mine are not in those locations. |
| Info. Request | Question (Prop.) | the middle one is the darkest of the 3? |
| | Question (Set) | where is it in relation to the large med grey? |
| | Question (Choice) | Which should we choose? / the black or the grey? |
| | Question (Check) | It's the darkest dot in the circle, right? |
| Commissives | Offer | lets click the upper left one that's bigger and darker gray |
| Directives | Request | tell me about your tiniest dot? / pick one at the bottom |
| | Suggestion | Please describe it in relation to other dots in the circle |

Table 3.3: Illustrative utterances in the dataset, grouped by the *task dimension* of communicative functions (Bunt et al., 2017)

vious mention (Poesio et al., 2016). This can occur both within utterances ("a medium size black one, with a very light slightly smaller one to *its* left") and across utterances ("Does *the lighter dot* appear to be slightly larger?"). In *associative anaphora*, the referents of the postcedent can be different from the antecedent, e.g. only refer to a *part of* the antecedent (as in "I have a pair where *the left one* is large and dark").

These discourse-level phenomena play a fundamental role in the process of common grounding and will be the main focus of our study in Chapter 4.

## 3.5 Experiments

### 3.5.1 Evaluation

In this experiment, we focus on a simple subtask of common grounding, which we refer to as the **target selection task**. Specifically, our goal is to predict which target entity a human player selected, provided the player's private observation and the corresponding dialogue. This is an essential subtask of collaborative reference, where the player makes the final selection based on the established common ground.

Since the number of entities in each player's view is fixed at 7, we can formulate this as a simple classification problem. However, we expect that even the accurate recognition of the target (i.e. common ground) would be challenging due to the complexity of common grounding strategies (as we discussed in Section 3.4).

### 3.5.2 Model Architecture

As a preliminary experiment, our baseline models are kept as simple as possible with minimal preprocessing and hyperparameter tuning. The two main components of the models are as follows:

#### Context Encoding

The dialogue context (agent's view) is represented as a 28-dimensional real-valued vector, where each of the 7 observable entities is represented as a 4-dimensional vector (x-value, y-value, size, color). As a preprocessing step, we normalize each dimension of the vector in the range of $(-1, 1)$.

The simplest way to encode this is to directly apply a multi-layered perceptron (MLP) over the context vector. However, without feature engineering, this simple approach may have difficultly capturing relevant information, such as the *relations* between entities. Therefore, in the second approach, we use the Relation Network (Santoro et al., 2017) to create additional features for the relations between entities. Specifically, we encode each *pair* of the entities with a shared MLP (for a total of 21 pairs) and append the sum of these vectors as an additional input.

**Dialogue Encoding**

Utterances are all tokenized and lowercased, and tokens which occur less than 10 times are treated as a unique *unknown* token. We insert a token which represents the *speaker id* to each utterance at the beginning, and another token to indicate the end of the dialogue. Then, we embed each token with a shared MLP and run a bidirectional GRU (Cho et al., 2014) over the embedded tokens. Finally, we take the last state of the bi-GRU as the final representation (encoding) of the dialogue.

For prediction, we simply concatenate the context and dialogue encodings and run another MLP. However, as we've seen in Section 3.4, there exists nonlinguistic selection bias in our dataset which makes predictions possible without using linguistic information (i.e. dialogue encoding). Therefore, as an ablation study, we also train the models to make predictions based on the context encodings only.

Following common practice, we split the dataset into training, validation and test set with a proportion of 8:1:1, and all models are tuned on the validation set. The loss function is calculated using cross entropy. All components of the neural networks are single-layered with 128 hidden units, and a dropout rate of 0.5 is applied at each layer to avoid overfitting. All parameters are initialized uniformly within the range of $(-0.01, 0.01)$. Models are trained with the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 0.001, and we clip gradients whose $L^2$ norm is greater than 0.1. The experiment is run 10 times initialized with different seeds, and we report the mean and standard deviation of the selection accuracies on the full test set.

For further analyses, models with the best validation loss in the previous experiment are also evaluated on two variants of the test set. First, since the current test set contains correlated predictions on the same dialogue from the two players, we create the *uncorrelated* test set by randomly removing one of the player's prediction. Secondly, we further removed dialogues where players failed to coordinate on the same entity to create the *success-only* test set. The statistical significance of the results are computed on the uncorrelated test set using paired student's t-test.

Finally, we take 100 random samples from the uncorrelated test set (including 76 successful dialogues) to report human performance based on the average accuracy of two annotators.

### 3.5.3 Results

We show the results of our experiment in Table 3.4. As we can see, models trained with the context encodings only perform significantly better than random ($p$-value $< 10^{-7}$). This verifies that we must indeed take into account the effect of selection bias when interpreting model performances.

Secondly, we found that encoding context with Relation Network (RN) consistently outperforms MLP, but not at a statistically significant level ($p$-value $> 0.1$). Therefore, the simplest strategy of using MLP works decently, but an improved architecture may potentially improve the overall performance.

| | Full | Uncorrelated | Success Only |
|---|---|---|---|
| Random | 14.28 | 14.28 | 14.28 |
| Context Only (MLP) | $27.90 \pm 0.6$ | 28.74 | 29.59 |
| Context Only (MLP+RN) | $31.94 \pm 0.9$ | 30.22 | 32.40 |
| Context + Dialogue (MLP) | $40.27 \pm 1.3$ | 40.89 | 43.82 |
| Context + Dialogue (MLP+RN) | $43.09 \pm 0.8$ | 44.00 | 49.44 |
| Human | - | 82.50 | 90.79 |

Table 3.4: Results of the target selection task.

Thirdly, models trained with both context and dialogue encodings significantly outperform models trained only with the context encodings ($p$-value $< 10^{-9}$). This indicates that even our simplest models can learn to leverage linguistic information along with the context to make better predictions.

Finally, we found that when the test set only includes successful cases, both the models and humans perform consistently better: however, human performance improves even more dramatically, achieving over 90% accuracy. This indicates that the *success-only* test set contains higher-quality dialogues with less noise.

Overall, our target selection task is challenging due to the complexity of common grounding, and we still have a huge room left for improvement.

## 3.6 Conclusion

In this study, we proposed a novel task setting under continuous and partially-observable context to require advanced strategies of common grounding. Based on this task setting, we formulated a minimal collaborative reference task to measure the ability of creating accurate common ground. To enable various empirical studies, we collected a large-scale dataset of actual human dialogues (OneCommon Corpus) through a careful crowdsourcing procedure. Our dataset analyses revealed the difficulties and distinct strategies of common grounding involved in our task. Finally, we developed and analyzed simplest baseline models based on the subtask of recognizing common ground. Due to the complexity of common grounding, we showed that there remains major room left for improvement in future work.

# Chapter 4

# Interpretation Based on Reference Resolution

Interpreting the process of common grounding is a challenging task, especially under continuous and partially-observable context where complex ambiguity, uncertainty, partial understandings and misunderstandings are introduced. Interpretation becomes even more challenging when we deal with existing dialogue systems which still have limited capability of natural language understanding and generation. To address this problem, we consider reference resolution as the central subtask of common grounding and propose a new resource to study its intermediate process. Based on a simple and general annotation schema, we collected a total of 40,172 referring expressions in 5,191 dialogues curated from OneCommon Corpus, along with multiple judgements of referent interpretations. We show that our annotation is highly reliable, captures the complexity of common grounding through a natural degree of reasonable disagreements, and allows for more detailed and quantitative analyses of common grounding strategies. Finally, we demonstrate the advantages of our annotation for interpreting, analyzing and improving common grounding in baseline end-to-end dialogue systems.

## 4.1 Introduction

Common grounding is the process of creating, repairing and updating mutual understandings, which is a critical aspect of sophisticated human communication (Clark, 1996) as well as a longstanding goal in dialogue modeling (Traum, 1994). Recently, there have been new proposals of dialogue tasks which require advanced skills of common grounding under *continuous* and *partially-observable* context (Udagawa and Aizawa, 2019; Haber et al., 2019). Their main contributions include the establishment of clear evaluation metrics (based on task success rates), collection of large-scale datasets and introduction of complex ambiguity, uncertainty, partial understandings and misunderstandings which are minimally observed under traditional (either categorical or fully-observable) context.

However, interpretation of the process of common grounding remains an open problem. Although formal approaches such as Lascarides and Asher (2009) and Poesio and Rieser (2010) account for some of the important details in common grounding, constructing such precise semantic representations is a difficult and costly process, especially under continuous and partially-observable context with high ambiguity and uncertainty. Interpretation becomes even more challenging when we deal with existing dialogue systems represented by the end-to-end dialogue systems (Vinyals and Le, 2015; Bordes et al., 2017), which can converse fluently but still lack the true competence of natural language understanding and generation.

In this study, we approach this problem by *decomposing* the common grounding task based on its intermediate subtasks. Specifically, we consider *reference resolution* as the central subtask of common grounding (in the sense that mutual understanding can only

be created through successful references), define this subtask formally based on a simple and general annotation schema, and create a large-scale resource to study this subtask along with the original task of common grounding.

Our annotated corpus consists of a total of 40,172 referring expressions in 5,191 dialogues curated from OneCommon Corpus (Udagawa and Aizawa, 2019), along with multiple (a minimum of 3) judgements for referent interpretations. A visualization of our annotation is shown in Figure 4.1.

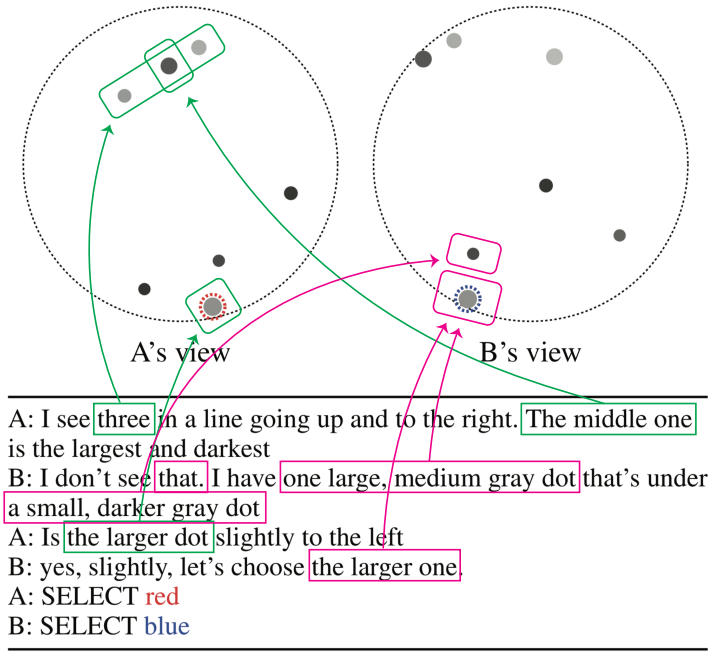



Figure 4.1: A visualized example of our annotation. We identify all referring expressions in the dialogue and their intended referents based on the speaker's perspective (only one judgement shown in this example).

Through our corpus analyses, we show that our annotation has high agreement in general but also includes a natural degree of reasonable disagreements, which verified that our annotation can be conducted reliably while capturing the ambiguity and uncertainty under continuous and partially-observable context. In addition, we give a more quantitative analysis of *pragmatic expressions* as an illustrative example of analyses that can be conducted based on our annotation.

Finally, in our experiments, we show that our annotation is critical for interpreting and analyzing common grounding in baseline end-to-end dialogue systems, as well as improving their performance on difficult end tasks.

Overall, the contributions of this chapter are as follows:

- We proposed a novel method of decomposing common grounding based on the central subtask of *reference resolution* to study the intermediate process of common grounding.
- We conducted a large-scale annotation of 5,191 dialogues from OneCommon Corpus, including 40,172 referring expressions with multiple judgements for referent identification.
- We verified the *reliability* of our annotation (while capturing genuine ambiguity based on reasonable disagreements) as well as the *usefulness* for analyzing human common grounding strategies.

- Our experiments demonstrate that our annotation can be utilized for both interpreting and improving common grounding in end-to-end dialogue systems.

## 4.2 Annotation Procedure

The goal of our annotation is to provide a *general*, *reliable* and *useful* annotation of reference resolution to study the intermediate process of common grounding. In this work, we use the 5,191 dialogues in OneCommon Corpus which succeeded on the collaborative reference task, since they are expected to be of higher quality (c.f. Section 3.5.3). Our annotation procedure consists of two main steps: *markable detection* to semi-automatically detect the referring expressions under consideration, followed by *referent identification* to identify their referents.

As an optional step, we also conducted *preprocessing* of the dialogues to correct obvious misspellings and grammatical errors. Due to the limited size of the vocabulary, we manually looked for rare unigrams and bigrams in the dialogue and carefully designed rules to correct them. Our preprocessing step is reversible, so the collected annotation can also be applied to the original dialogues without preprocessing.

### 4.2.1 Step 1: Markable Detection

In this study, we define a *markable* to be an independent referring expression of the entities under consideration: in our case, the synthetic entities on the 2-D plane. Basically, we annotate a markable as a minimal noun phrase including all prenominal modifiers (such as determiners, quantifiers, and adjectives) but excluding all postnominal modifiers (such as prepositional phrases and relational clauses). This eliminates the complexity of the annotation because markables will not overlap or nest with each other.

To reduce the annotation effort in the later process, we optionally annotate three attributes for each markable if they are obvious from the context: a *generic* attribute when the markable is not specific enough to identify the referents, *all-referents* when the markable is referring to all of the entities in the speaker's view, and *no-referent* when the referents are empty. *Generic* markables are ignored in our annotation, and the referents of *all-referents* or *no-referent* are annotated automatically in the later process. To reduce the redundancy of annotation, we consider a predicative noun phrase as a markable only if there is no previous markable in the same utterance that refers to the same entities: for example, "*a triangle*" in "three dots are forming *a triangle*" is not considered as a markable since "three dots" is already annotated, but it is considered a markable in "one light dot and two dark dots are forming *a triangle*" (underlines indicate markables). We also annotate obvious *anaphoric* and *cataphoric* relations in the same utterance if they have identical referents: this way, the referents of anaphoric/cataphoric markables can be annotated automatically based on their antecedents/postcedents. However, we do not annotate such relations *across utterances* as they can potentially refer to different entities, e.g. in the case of misunderstandings (as shown in Figure 4.2).

The annotators were graduate students with sufficient experience and training, and we used the brat annotation tool (Stenetorp et al., 2012) to detect the markables, their attributes and relations. All available information were accessible during the annotation, including the original dialogues, players' observations and final selections.

### 4.2.2 Step 2: Referent Identification

Next, we used crowdsourcing on Amazon Mechanical Turk to collect large-scale judgements of the referents for each markable. Our visual interface for referent identification is shown in Figure 4.3. Annotators were instructed to read the instructions carefully,

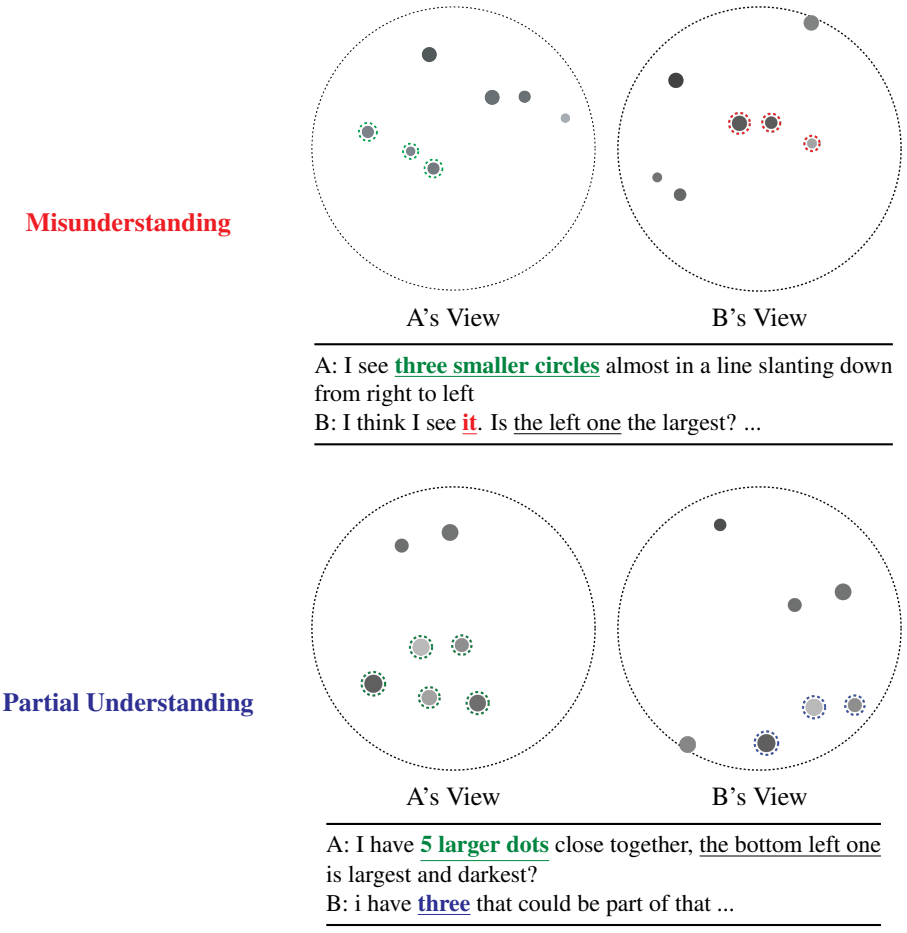



Figure 4.2: Illustrative examples of misunderstanding and partial understanding captured by our annotation.

including the description of the underlying collaborative reference task. If the referents were ambiguous, they put a check on *ambiguous* box and selected all possible candidate referents. If the referents were completely unidentifiable based on the available information, they put a check on *unidentifiable* box (without selecting the referents).

To collect reliable annotations, we restriceted the workers to those with at least 100 previously completed HITs (submissions on AMT) and above 99% acceptance rate. We paid the workers well, with $0.25 for dialogues with less than 7 markables, $0.35 with less than 14 markables, and $0.45 otherwise. In addition, we automatically detected outlier submissions based on several statistics (such as the agreement with other workers) and manually reviewed them to encourage better work or reject clearly unacceptable works. The overall rejection rate was 1.18%.

As a result of this careful crowdsourcing, we were able to collect a large-scale annotation of 103,894 judgements with at least 3 judgements for each of the 34,341 markables that required manual referent identification. As shown in Figure 4.2, our annotation captures important phenomena of the intermediate process of common grounding (e.g. misunderstandings and partial understandings).

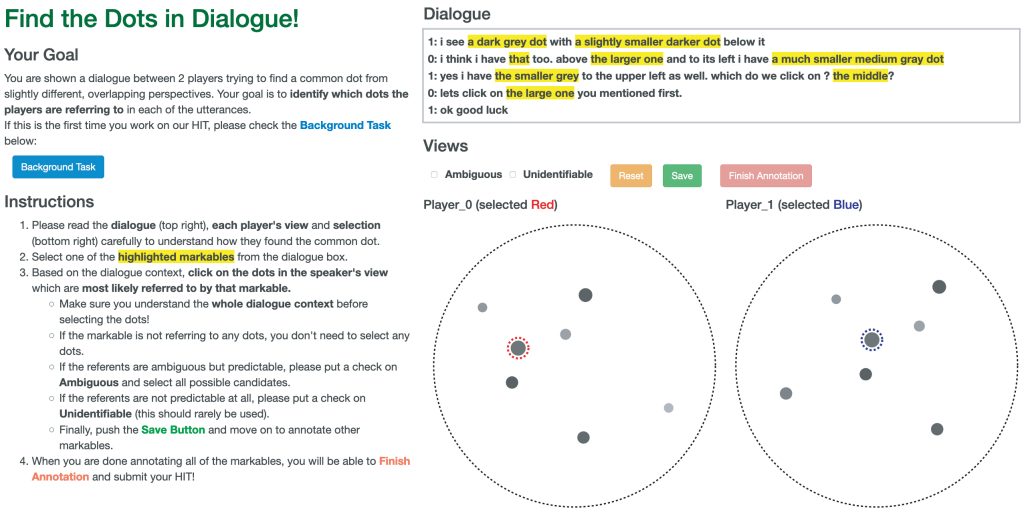



Figure 4.3: Our visual interface used for referent identification.

| # Markables | # All-Referents | # No-Referent | # Anaphora | # Cataphora | Agreement (%) | |
|---|---|---|---|---|---|---|
| | | | | | Start Pos. | End Pos. |
| 40,172 | 128 | 1,149 | 4,548 | 6 | 99.11 (96.32) | 99.06 (96.11) |

Table 4.1: Annotation statistics for markable detection. Agreement is calculated at the token level (Fleiss's Multi-$\pi$ shown in parenthesis).

| # Markables | # Judgements | Ambiguous (%) | Unidentifiable (%) | Agreement (%) | Exact Match (%) |
|---|---|---|---|---|---|
| 34,341 | 103,894 | 4.65 | 0.77 | 96.26 (88.66) | 86.90 |

Table 4.2: Annotation statistics for referent identification, along with the rate of *ambiguous* and *unidentifiable* checked in the judgements. Agreement is calculated at the entity-level (Fleiss's Multi-$\pi$ in parenthesis) and exact match rate at the markable-level.

## 4.3 Annotation Results

### 4.3.1 Annotation Statistics

First, we report the annotation statistics of markable detection in Table 4.1 and referent identification in Table 4.2. All agreements are computed based on pairwise judgements, and we use Fleiss's Multi-$\pi$ (Fleiss, 1971) to remove the effect of chance level agreements. For markable detection, agreement is calculated for the markable text span (start/end positions at the token level) based on 130 dialogues by 3 annotators. Agreements for markable attributes and relations were also reasonable, but we omit the results since they were optional and used only for the purpose of automatic referent identification. For referent identification, agreement is calculated at the entity-level (whether each entity is included in the referents) and the exact match rate at the markable-level (whether the set of referents match exactly).

Overall, we found high agreement for all annotations (including the crowdsourcing step), which verified the reliability of our annotation framework.

### 4.3.2 Disagreement Analysis

However, it is natural that there is a certain degree of disagreement in referent interpretations. In fact, it is important to capture such disagreements as there can be genuine

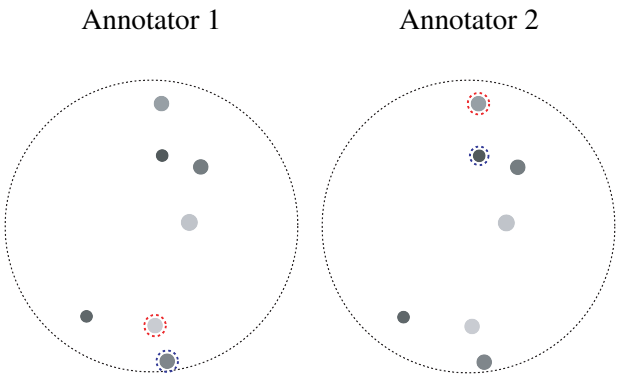



Figure 4.4: An example of seemingly reasonable disagreement captured by our annotation.

| # Referents | Agreement (%) | Exact Match (%) | Judgements (%) |
|---|---|---|---|
| 0 | 78.04 | 17.78 | 1.31 |
| 1 | 97.45 | 90.28 | 71.81 |
| 2 | 94.87 | 82.17 | 14.85 |
| 3 | 93.93 | 83.03 | 7.51 |
| 4 | 92.18 | 76.66 | 2.20 |
| 5 | 90.31 | 71.03 | 0.88 |
| 6 | 90.75 | 78.14 | 1.22 |
| 7 | 81.47 | 62.50 | 0.21 |

Table 4.3: Agreement statistics conditioned on the number of referents (and the percentages of such judgements).

ambiguity and uncertainty under continuous and partially-observable context (see Figure 4.4 for an example). Therefore, in addition to *explicitly* annotating the ambiguity and unidentifiability (as described in Section 4.2.2), we aim to capture them *implicitly* by collecting multiple judgements from different annotators.

To study such disagreements in detail, we computed the agreement statistics (e.g. entity-level agreement) conditioned on *the number of referents* in each judgement. To be specific, given a number of referents (from 0 to 7), we collected all judgements with that number of referents and computed the average pairwise agreements with all other judgements on the same markable. The results are summarized in Table 4.3.

We can see that there is a significant amount of disagreements when the number of referents was judged to be 0 (none) or 7 (all). This could be due to several reasons: obvious cases were already annotated as *no-referent* or *all-referents* during markable detection (so only difficult cases were left), annotators simply made mistakes (e.g. forgot to annotate), or the referents were annotated as such when it was too difficult to identify them (e.g. all entities were potential referents). Since the number of such judgements were relatively small, their effect can be mitigated after appropriate aggregation of multiple judgements. We also expect that they provide a useful resource for studying disagreements caused by either the *annotation error* or *genuine ambiguity*, which is a critical problem when multiple interpretations are possible (Poesio et al., 2019).

In addition, we found that the exact match rate is the highest when the referent is only 1 and much lower as the number of referents increases. This is reasonable because referring expressions of multiple entities tend to be more pragmatic and ambiguous (e.g. *"a cluster"*, *"most of"*, *"a line"*), and it would be more difficult to match the referents

exactly. Note that entity-level agreements are still at a high level, and the interpreted referents seem to mostly overlap with each other.

Finally, to study which expressions tend to have higher (or lower) disagreements, we computed the correlations between the occurrence of common tokens (in the markable text) and the exact match rate (of the pairwise judements for the markable). Illustrative examples are shown in Table 4.4.

| Low | $\rho$ | # Count | High | $\rho$ | Count |
|---|---|---|---|---|---|
| *it* | -0.149 | 12.7K | *lower* | 0.028 | 1.3K |
| *any* | -0.103 | 0.5K | *two* | 0.030 | 14.7K |
| *that* | -0.100 | 12.5K | *three* | 0.031 | 4.2K |
| *your* | -0.083 | 1.5K | *darkest* | 0.036 | 2.1K |
| *few* | -0.081 | 0.1K | *larger* | 0.039 | 7.7K |
| *what* | -0.081 | 0.4K | *middle* | 0.041 | 2.1K |
| *others* | -0.064 | 0.8K | *smallest* | 0.043 | 2.0K |
| *line* | -0.062 | 1.7K | *very* | 0.056 | 6.1K |
| *bunch* | -0.060 | 0.2K | *top* | 0.061 | 5.2K |
| *all* | -0.048 | 1.1K | *light* | 0.072 | 18.7K |
| *triangle* | -0.046 | 2.5K | *tiny* | 0.076 | 7.8K |
| *some* | -0.042 | 0.2K | *large* | 0.084 | 21.7K |
| *medium* | -0.041 | 12.5K | *the* | 0.125 | 55.0K |
| *another* | -0.039 | 1.4K | *one* | 0.136 | 57.1K |
| *and* | -0.029 | 1.7K | *black* | 0.145 | 26.9K |

Table 4.4: Tokens with low or high correlation with the exact match rate (Pearson's correlation coefficient shown in $\rho$).

In general, the correlations are very small and the amount of disagreements seem relatively constant across all token types. However, the general trend is still intuitive: ambiguous or complex expressions such as pronouns, interrogatives, quantifiers, and conjunctions tend to have negative correlations, while simple and plain expressions tend to have positive correlations.

To summarize the analysis, our annotation has high overall agreement but also includes interesting, reasonable disagreements which capture the ambiguity and uncertainty under continuous and partially-observable context.

### 4.3.3 Pragmatic Expressions

Finally, we give an illustrative example of additional analyses that can be conducted based on our annotation. To be specific, we give a more quantitative analysis of *pragmatic expressions* which we introduced as a characteristic strategy under continuous context (Section 3.4.1).

As an illustration, we focus on the pragmatic expressions of *color* and estimate the distributions of the actual color of the referents described by the common adjectives. We simply assume that the adjective in the markable (minimal noun phrase) describes the color of the referents, since the exceptions (such as negation in the prenominal modifier) seemed rare and ignorable. For the sake of visualization, empirical distributions are smoothed based on kernel density estimation. As we can see in Figure 4.5, all adjectives (including the specific color "black") have relatively wide distributions which overlap with each other. This is a strong evidence that the color expressions are pragmatic, i.e. the same adjective can be used for different colors (and in return, the same color can be described in different ways) depending on the context.

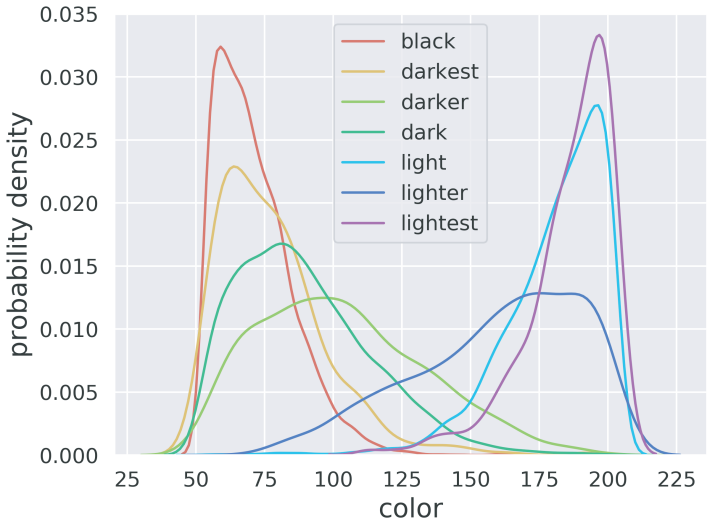



Figure 4.5: Distributions of the actual color of the referents expressed by common adjectives (the range of color is 256 in grayscale, lower is darker).

## 4.4 Experiments

In this experiment, we evaluate and analyze baseline dialogue models based on the following three tasks:

- First is the **target selection task** from Chapter 3, where the model predicts the target entity selected by each player at the end of collaborative reference. This task requires accurate recognition of the created common ground.
- Second is the **reference resolution task**, where the model predicts the referent entities for each markable in the dialogue. Note that the model is given only one player's observation and predicts the markables in that player's utterances only. This task requires accurate comprehension of the intermediate process of common grounding.
- Last is the **selfplay dialogue task**, where the model plays the whole collaborative reference task against an identical copy of itself. This requires the actual creation of common ground through natural language communication, despite against the copy of itself (and not with real humans).

To create the golden labels for reference resolution, we used simple majority voting based on the multiple judgements or automatically identified the referents based on markable attributes/relations. Note that markables were removed if the majority considered their referents as unidentifiable (in the referent identification step).

### 4.4.1 Model Architecture

The overall architecture of our baseline end-to-end models is shown in Figure 4.6.

#### Encoders

Our baseline models have two encoders: one for encoding dialogue tokens and one for context information.

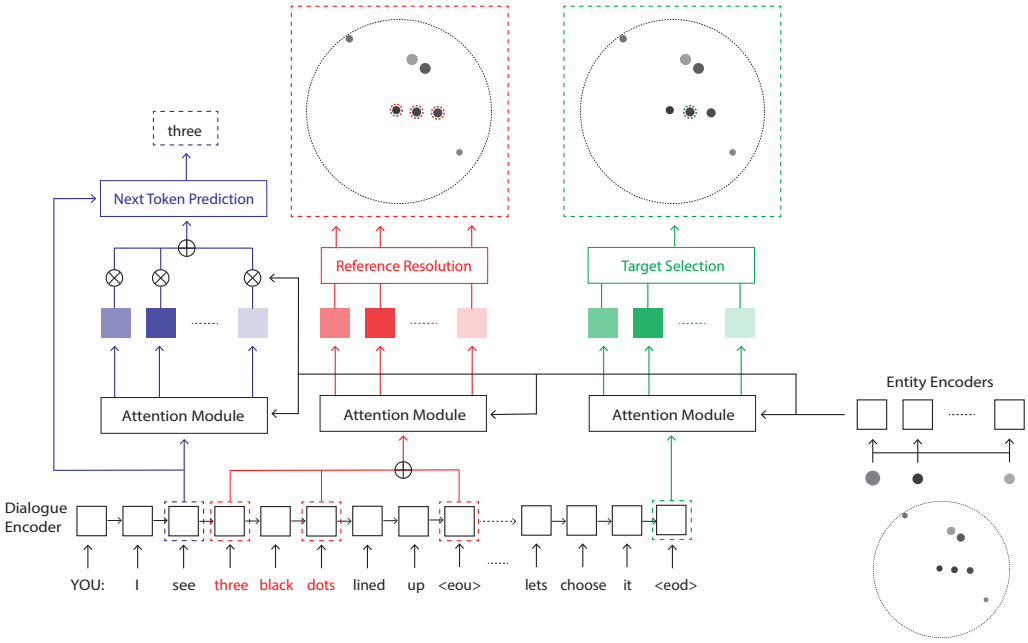



Figure 4.6: Our baseline model architecture (best seen in color). TSEL decoder is shown in green, REF decoder and the input markable ("three black dots") are in red, and DIAL decoder is in blue. All decoders share (some or all layers of) the attention module.

Dialogue tokens are encoded based on a unidirectional GRU (Cho et al., 2014). To encode context information, we use a shared *entity encoder* to create entity-level representations of the context. This consists of two modules: an *attribute encoder* which encodes the attributes of each entity (size, color and 2-D location) with a matrix followed by a tanh layer, and a *relational encoder* which encodes relative attributes of each entity pairs with another matrix followed by a tanh layer. The final representation of each entity is the concatenation of its attribute encoding and the sum of relational encodings with the other 6 entities. Note that we refined the method from Chapter 3, where the context is encoded into a single representation (and not entity-level representations).

### Decoders

Our models can have up to three decoders: TSEL for target selection, REF for reference resolution, and DIAL for predicting next dialogue tokens. Each decoder shares (some or all layers of) the *attention module* based on an MLP which computes a scalar score for each entity. Specifically, this module takes the representation of each entity and the following positions (states) of the GRU as the input: the *final hidden state* for TSEL, the *start/end positions of the markable* and the *end position of the utterance* for REF, and the *current hidden state* for DIAL. Based on the computed attention scores, TSEL simply computes the softmax and REF computes logistic regressions for each entity. DIAL reweights the entity representations based on their attention scores to focus on the relevant entities in next token prediction (Bahdanau et al., 2014; Xu et al., 2015).

In this experiment, we built five models based on different combinations of the three decoders. All models are trained with the same default hyperparameters, following the similar setup as Chapter 3.

| Model | Target Selection | Reference Resolution (Accuracy / Exact Match) | Selfplay Dialogue #Shared=4 | #Shared=5 | #Shared=6 |
|---|---|---|---|---|---|
| TSEL | 67.79±1.53 | - | - | - | - |
| REF | - | **85.75±0.22 / 33.91±0.86** | - | - | - |
| TSEL-REF | **69.01±1.58** | 85.47±0.36 / 32.88±1.28 | - | - | - |
| TSEL-DIAL | 67.01±1.29 | - | 42.07±1.27 | 57.37±1.29 | 77.00±1.13 |
| TSEL-REF-DIAL | **69.09±1.12** | 85.86±0.18 / 33.66±0.93 | **45.78±2.15** | **61.95±1.72** | **80.01±1.61** |
| Human | 90.79 | 96.26 / 86.90 | 65.83 | 76.96 | 87.00 |

Table 4.5: Results of our experiments. For reference resolution, *accuracy* is computed at the entity-level and *exact match rate* at the markable-level. Human scores are taken from Table 3.1 and 4.2 as a reference.

### 4.4.2 Results

We run the experiments 10 times with different random seeds and dataset splits. For selfplay dialogues, we generated 1,000 scenarios with each number of shared entities (4, 5 or 6) and set the output temperature to 0.25 during next token prediction. We report the mean and standard deviation of the results in Table 4.5.

In terms of *target selection* and *selfplay dialogue* tasks, we found consistent improvements by training the models jointly with reference resolution (i.e. with REF decoder). This verified that even simple multi-task training with the central subtask can improve performance on difficult end tasks. The results for *reference resolution* are reasonably high in terms of entity-level accuracy but much lower in terms of exact match rate. Considering the high agreements (Section 4.3.1) and improved reliability of the gold annotation after aggregation, we expect there to be a huge room for further improvements.

Overall, common grounding under continuous and partially-observable context is still a challenging task, and we expect our resource to be fundamental for solving this task along with the accurate capability of reference resolution.

### 4.4.3 Further Analysis

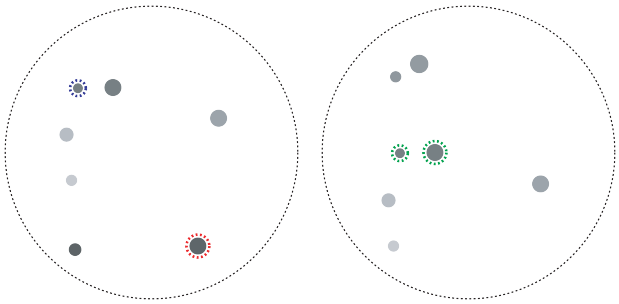
To demonstrate the advantages of our annotation for interpreting and analyzing dialogue systems, we give a more detailed analysis of TSEL-REF-DIAL model which performed well on all three tasks. In Table 4.6, we show the results for reference resolution grouped by the number of referents in the gold annotation. In terms of the exact match rate, we found that the model performs very well on 0 and 7 referents: this is because most of them can be recognized at the superficial level, such as “none of them", “all of mine", “I don’t have that", etc. However, the model struggles on all other cases: the results are especially worse for markables with more than one referent. This shows that the model still lacks the ability of precisely tracking multiple referents, which can be expressed in complex, pragmatic ways (based on the grouping strategies).

In addition, we found that the correlation between the reference resolution accuracy (i.e. average accuracy of reference resolution in each dialogue) and the target selection accuracy (i.e. binary result of target selection in each dialogue) was relatively weak, with an average of only 0.23 in the 10 runs of the experiments. This reveals that the model is often correct for the target selection task based on the *wrong reason*, without tracking the referents correctly (McCoy et al., 2019). Our annotation is also useful for error analyses in recognizing common ground, e.g. by inspecting *where* the model made a mistake and lost track of the correct referents.

Finally, we show an example dialogue from the selfplay task with the interpreted process of common grounding (Figure 4.7). Referring expressions are automatically detected by a BiLSTM-CRF tagger (Huang et al., 2015) trained on our corpus with the

| # Referents | Accuracy (%) | Exact Match (%) | # Count |
|---|---|---|---|
| 0 | 95.91±1.38 | 83.53±4.65 | 148.5 |
| 1 | 89.34±0.17 | 36.86±1.32 | 2782.5 |
| 2 | 78.14±1.07 | 20.59±1.90 | 587.9 |
| 3 | 70.64±1.02 | 13.63±2.06 | 283.3 |
| 4 | 69.12±2.69 | 10.16±3.47 | 81.0 |
| 5 | 73.57±2.94 | 17.56±5.88 | 33.0 |
| 6 | 78.69±4.45 | 13.18±7.31 | 43.0 |
| 7 | 74.60±7.49 | 50.38±11.40 | 22.3 |

Table 4.6: Detailed results for the reference resolution task grouped by the number of referents in the gold annotation (along with the average counts in the test set).



Model A's view Model B's view

A: I have **a large black dot** with **a smaller dark dot** to the right of **it**
B: I see **that** . Let's pick the large black dot

Figure 4.7: An example dialogue from the selfplay dialogue task by the TSEL-REF-DIAL model. Predicted referents are highlighted (no referents are predicted for "the large black dot").

result of 98.9% accuracy at the token level. Based on the raw dialogue only, it is difficult to identify which entities model A is referring to. However, by visualizing the intended referents, we can see that model A is describing two entities in somewhat unnatural and inappropriate way (albeit using the anaphoric expression "it" appropriately). We can also verify that model B acknowledges this in a perfectly coherent way but without predicting any referent for "the large black dot": we often observed such phenomena, where the utterance by a model cannot be interpreted correctly *even by itself*. Overall, our annotation allows for fine-grained analysis of both the *capabilities* and *incapabilities* of existing dialogue systems. The generated dialogue is short in this example, but our approach would be even more critical for interpretation as the dialogues get longer and more complicated.

## 4.5 Related Work

Coreference and anaphora resolution have been studied extensively in NLP (Pradhan et al., 2011; Poesio et al., 2016), including disagreements in their interpretations (Recasens et al., 2012; Poesio et al., 2019). The main novelty of our annotation schema is the focus on *exophoric references* and direct annotation of the referents in the *external modality*. This allows us to study the aspect of *symbol grounding* in the visual context, as we discuss in the next chapter. This also enables reliable and intuitive annotation, even by using non-expert annotators for referent identifications. Finally, our annotation (at least indirectly) captures basic coreferences as well as complex associative anaphora

(such as *part-of* relations): with additional annotations, we can also capture more fine-grained, explicit relations between anaphora as well.

Our work is also relevant to the recent literature of interpretable and explainable machine learning (Doshi-Velez and Kim, 2017; Lipton, 2016). Especially the analysis of neural based models is gaining attention in NLP (Belinkov and Glass, 2019), including end-to-end dialogue models (Sankar et al., 2019). The main novelty of our approach is that we decompose the original task (*common grounding*) based on its central subtask (or could be subtasks), define the subtask (*reference resolution*) formally with an annotation framework, and create a large-scale resource to study the subtask along with the original task. Our approach has several advantages compared to previous analysis methods. First, it is applicable to *both humans and machines*, which is especially important in dialogue domains where they interact. Second, it can be used to study the *relationships* between the original task and its subtasks, which is critical for a more *skill-oriented* evaluation of artificial intelligence (Hernndez-Orallo, 2017; Sugawara et al., 2017). Third, it can be used for investigating *the dataset* on which the models are trained: this is important in many aspects, such as understanding undesirable biases in the dataset (Gururangan et al., 2018; Sugawara et al., 2018) or correct model predictions based on the *wrong reasons* (McCoy et al., 2019). Finally, the collected resource can be used for both *probing* whether the models solve the subtasks implicitly (Linzen et al., 2016) or *developing* new models which can be explicitly supervised, evaluated and interpreted based on the subtasks.

Finally, visually grounded dialogues have been studied in a wide variety of settings. In comparison, the main strengths and novelty of our (annotated) OneCommon Corpus can be summarized as follows:

(A) Our corpus is based on the advanced setting of continuous and partially-observable context where complex common grounding strategies are introduced.

(B) Our corpus has the simplicity and controllability to make fully controlled experiments and analyses possible.

(C) Our corpus contains large-scale manual annotation of reference resolution and detailed analyses of agreements/disagreements based on multiple judgements.

Prior work in common grounding (Potts, 2012; De Vries et al., 2017) and visual reference resolution (Tokunaga et al., 2012; Zarrieß et al., 2016) mostly focus on categorical or fully-observable settings and do not satisfy (A). While visual dialogues based on photographic scenes (Das et al., 2017a; Haber et al., 2019; Ilinykh et al., 2019) have the strengths of being more complex and realistic, they do not satisfy (B). Götze and Boye (2016) conducted a smaller-scale annotation of reference resolution but do not assess the reliability of the annotation, hence not satisfying (C). To the best of our knowledge, our work is the first (and only) resource to satisfy all of the above criteria.

## 4.6 Conclusion

One of the most influential models of common grounding to date is the *contribution* model of Clark and Schaefer (1989): however, applying such theory in realistic settings can be difficult or even problematic. In this chapter, we proposed a novel method of decomposing common grounding based on the subtask of reference resolution to study the intermediate process of common grounding. Based on our annotated corpus, we demonstrated the advantages of our approach for analyzing human strategies as well as interpreting and improving baseline (end-to-end) dialogue models. Overall, we expect our study to be a fundamental step towards interpreting and improving common grounding through reference resolution.

# Chapter 5

# Linguistic Analysis Based on Spatial Expressions

Recent models achieve promising results in visually grounded dialogues. However, existing datasets often contain undesirable biases and lack sophisticated linguistic analyses, which make it difficult to understand how well current models recognize their precise linguistic structures. To address this problem, we make two design choices: first, we focus on our OneCommon Corpus, which contains minimal bias by design. Second, we analyze their linguistic structures based on *spatial expressions* and provide comprehensive and reliable annotation for 600 dialogues. We show that our annotation captures important linguistic structures including predicate-argument structure, modification and ellipsis. In our experiments, we assess our improved baseline's understanding of these structures through reference resolution (Chapter 4). We demonstrate that our annotation reveals both the strengths and weaknesses of our baseline in essential levels of detail. Overall, we propose a novel framework and resource for investigating fine-grained language understanding in visually grounded dialogues.

## 5.1 Introduction

Visual dialogue is the task of holding natural, often goal-oriented conversation in a visual context (Das et al., 2017a; De Vries et al., 2017). This typically involves two types of advanced grounding: *symbol grounding* (Harnad, 1990), which bridges symbolic natural language and continuous visual perception, and *common grounding* (Clark, 1996), which refers to the process of developing mutual understandings through successive dialogues. As noted in Monroe et al. (2017) and Udagawa and Aizawa (2019), the *continuous* nature of visual context introduces challenging symbol grounding of nuanced and pragmatic expressions. Some further incorporate *partial observability* where the agents do not share the same context, which introduces complex misunderstandings and partial understandings that need to be resolved through advanced common grounding (Udagawa and Aizawa, 2019; Haber et al., 2019).

Despite the recent progress on these tasks, it remains unclear what types of linguistic structures can (or cannot) be properly recognized by existing models for two reasons. First, existing datasets often contain undesirable biases which make it possible to make correct predictions *without* recognizing the precise linguistic structures (Goyal et al., 2017; Cirik et al., 2018; Agarwal et al., 2020). Second, existing datasets severely lack in terms of sophisticated linguistic analysis, which makes it difficult to understand what types of linguistic structures exist or how they affect model performance.

To address this problem, we make the following design choices in this work:

- We focus on our OneCommon Corpus, a simple yet challenging collaborative reference task under continuous and partially-observable context. In this dataset, the visual contexts are kept simple and controllable to remove undesirable biases

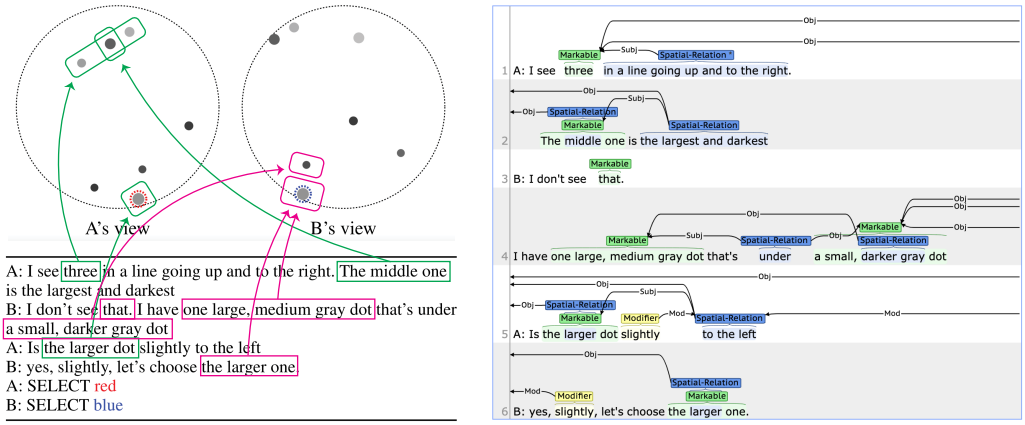



Figure 5.1: Example dialogue from OneCommon Corpus with reference resolution annotation (left) and our spatial expression annotation (right). We consider spatial expressions as predicates and annotate their arguments as well as modifiers.

while enhancing linguistic variety. In total, 5,191 successful dialogues are fully annotated with referring expressions (called *markables*) and their referents, which can be leveraged for further linguistic analysis.

- To capture the linguistic structures in these dialogues, we propose to annotate *spatial expressions* which play a central role in visually grounded dialogues. We take inspiration from the existing annotation frameworks (Pustejovsky et al., 2011a,b; Petruck and Ellsworth, 2018; Ulinski et al., 2019) but also make several simplifications and modifications to improve coverage, efficiency and reliability.[1]

As shown in Figure 5.1, we consider spatial expressions as *predicates* with existing markables as their *arguments*. We distinguish the argument roles based on *subjects* and *objects*[2] and annotate *modifications* based on nuanced expressions (such as "slightly"). By allowing the arguments to be in previous utterances, our annotation also captures *argument ellipsis* in a natural way.

In our experiments, we focus on reference resolution to study the model's comprehension of these linguistic structures. Since we found the existing baseline to perform relatively poorly (especially on the exact match rate), we propose a simple method of incorporating *numerical constraints* in model predictions, which significantly improved its prediction quality.

Based on our annotation, we conduct a series of analyses to investigate whether the model predictions are *consistent* with the spatial expressions. Our main finding is that the model is adept at recognizing entity-level attributes (such as color and size) but mostly fails in capturing inter-entity relations (especially placements): using the terminologies from Landau and Jackendoff (1993), the model can recognize the *what* but not the *where* in spatial language. We also conduct further analyses to investigate the effect of other linguistic factors.

Overall, the contributions of this chapter are as follows:

- We proposed a novel framework of annotating spatial expressions by leveraging referring expressions in visual dialogues.

[1]For instance, we define *spatial expressions* in a broad sense and include spatial attributes (e.g. object size and color) as well as their comparisons.

[2]Our *subject-object* distinction corresponds to other terminologies such as *trajector-landmark* or *figure-ground*.

- We sampled 600 random dialogues from OneCommon Corpus with reference resolution annotation (c.f. Chapter 4) and further conducted the annotation of spatial expressions.
- Our annotation captures important linguistic structures in visual dialogues, including predicate-argument structure, modification and ellipsis.
- Based on our improved baseline, we assess how well the end-to-end models can recognize the precise linguistic structures in the reference resolution task.

## 5.2 Annotation Procedure

In this work, we randomly sample 600 dialogues from the annotated OneCommon Corpus (5,191 dialogues annotated with reference resolution) to conduct further annotation of spatial expressions. Our annotation procedure consists of three steps: *spatial expression detection*, *argument identification* and *canonicalization*.

### 5.2.1 Step 1: Spatial Expression Detection

Based on Pustejovsky et al. (2011a,b), spatial expressions are defined as the "constructions that make explicit reference to the spatial attributes of an object or spatial relations between objects".[3] We generally follow this definition and detect all spans of spatial attributes and relations in the dialogue. To make the distinction clear, we consider entity-level information (like color and size) as spatial attributes and other information (such as location and *explicit* attribute comparison) as spatial relations. Spatial attributes could be annotated as adjectives ("dark"), prepositional phrases ("of light color") or noun phrases ("a black dot"), while spatial relations could be adjectives ("lighter"), prepositions ("near"), and so on. We also detect the modification of spatial expressions based on negation and nuanced expressions, i.e. degree modifiers (c.f. Section 3.4.1).

Although we allow certain flexibility in determining their spans, holistic/interdependent expressions (such as "all shades of gray", "sloping up to the right", "very slightly") were instructed to be annotated as a single span. Independent expressions (e.g. connected by conjunctions) could be annotated separately or jointly if they had the same structure (e.g. same arguments and modifiers).

For the sake of efficiency, we do not annotate spatial attributes and their modifiers inside markables (see Figure 5.1), since their spans and arguments are easy to be detected automatically.

### 5.2.2 Step 2: Argument Identification

Secondly, we consider the detected spatial expressions as *predicates* and annotate referring expressions (markables) as their *arguments*. This approach has several advantages: first, it has broad coverage since referring expressions are prevalent in visual dialogues. In addition, by leveraging *exophoric* references which directly bridge natural language and the visual context, we can conduct essential analyses related to symbol grounding across the two modalities (c.f. Section 2.3).

To be specific, we distinguish the argument roles based on subjects and objects. We allow arguments to be in previous utterances *only if* they are unavailable in the present utterance. Multiple markables can be annotated for the subject/object roles, and no object need to be annotated in cases of spatial attributes, nominal/verbal expressions ("triangle", "clustered") or *implicit global objects* as in superlatives ("darkest (of

[3] Note that their term *object* corresponds to our term *entity*.

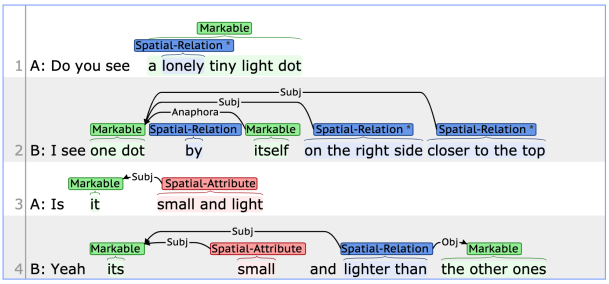



Figure 5.2: An example with spatial attributes (e.g. “small and light”).

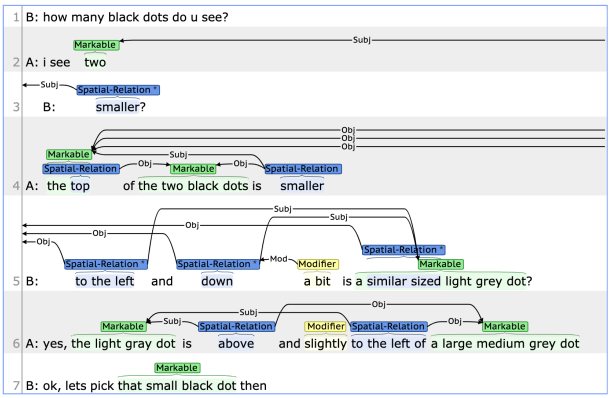



Figure 5.3: An example with subject ellipsis (“B: smaller?”).

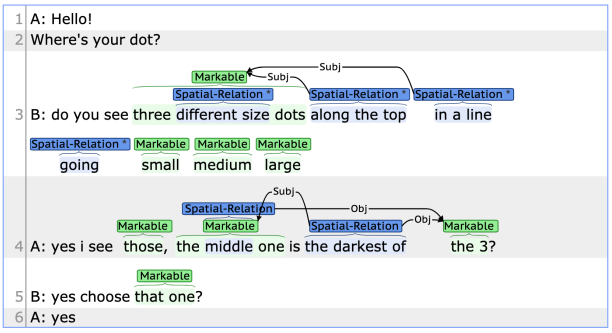



Figure 5.4: An example with unannotatable relation, “going (small medium large)”.

all)”). If the arguments are indeterminable based on these roles (as in enumeration, e.g. “From left to right, there are ...”), they were marked as *unannotatable*. Modificands of the modifiers (which could be either spatial attributes or relations) were also identified in this step. We show some illustrative examples in Figures 5.2, 5.3 and 5.4.

### 5.2.3 Step 3: Canonicalization

Finally, we conduct canonicalization of the spatial expressions and modifiers. Since developing a complete ontology for this domain is infeasible or too expensive, we focus on canonicalizing the central *spatial relations* in this work: we do not canonicalize spatial attributes manually since this can be conducted automatically (c.f. Section 5.5.1).

According to Landau (2017), there are 2 classes of relations in spatial language: *functional* class whose core meanings engage force-dynamic relationship (such as *on*, *in*) and *geometric* class whose core meanings engage geometry (such as *left*, *above*). Since functional relations are less common in this dataset and more difficult to define

due to their vagueness and context dependence (Platonov and Schubert, 2018), we focus on the following 5 categories of geometric relations and attribute comparisons, including a total of 24 canonical relations which can be defined explicitly.

- **Direction** requires the subjects and objects to be placed in a certain orientation: *left*, *right*, *above*, *below*, *horizontal*, *vertical*, *diagonal*.
- **Proximity** is related to the distance between subjects, objects or other entities: *near*, *far*, *alone*.
- **Region** restricts the subjects to be in a certain region determined by the objects: *interior*, *exterior*.
- **Color comparison** is related to the comparison of color among subjects and objects: *lighter*, *lightest*, *darker*, *darkest*, *same color*, *different color*.
- **Size comparison** is related to the comparison of size among subjects and objects: *smaller*, *smallest*, *larger*, *largest*, *same size*, *different size*.

To be specific, we annotate whether each detected spatial relation *implies* any of the 24 canonical relations. Each spatial relation can imply multiple canonical relations (e.g. "on the upper right" implies *right* and *above*) or none (e.g. "(forming a) triangle" does not imply any of the above relations).

In addition, we consider 6 modification types (the 5 degree modifiers from Figure 3.4 and *negation*) and canonicalize each modifier into one type. For example, "very slightly" is considered to have the overall type of the *diminisher*.

## 5.3 Annotation Results

### 5.3.1 Annotation Reliability

| Annotation | | Agreement (%) | Cohen's $\kappa$ |
|---|---|---|---|
| Span Detection | Attribute | 98.5 | 0.88 |
| | Relation | 95.1 | 0.87 |
| | Modifier | 99.2 | 0.86 |
| Argument Identification | Subject | 98.8 | 0.96 |
| | Object | 95.9 | 0.79 |
| | Modificand | 99.6 | 0.98 |
| Canonicalization | Relation | 99.7 | 0.96 |
| | Modifier | 87.5 | 0.83 |

Table 5.1: Results of our reliability analysis.

To test the reliability of our annotation, two trained annotators (the authors) independently detected the spatial expressions and modifiers in 50 dialogues. Then, using the 50 dialogues from one of the annotators, two annotators independently conducted argument identification and canonicalization. We show the observed agreement and Cohen's $\kappa$ (Cohen, 1968) in Table 5.1.

For span detection, we computed the token level agreement of spatial expressions and modifiers. Despite having certain freedom for determining their spans, we observed high agreement in general (including their start/end positions).

For argument identification, we computed the exact match rate of the arguments and modificands. As a result, we observed near perfect agreement for subject/modificand identification. For object identification, the result was comparatively worse: however, upon further inspection, we verified that 73.5% of the disagreements were essentially based on the same markables (e.g. coreferences).

Finally, we observed reasonably high agreement for relation/modifier canonicalization as well. Overall, we conclude that all steps of our annotation can be conducted with high reliability.

### 5.3.2 Annotation Statistics

| | Attribute | Relation |
|---|---|---|
| Total annotated | 378 | 4,300 |
| Average per dialogue | 0.63 | 7.17 |
| Unique expressions | 121 | 1,139 |
| Inter-utterance subject (%) | 1.59 | 1.37 |
| Inter-utterance object (%) | - | 14.65 |
| No object (%) | - | 30.84 |
| Unannotatable (%) | 0.79 | 0.79 |
| Modified (%) | 36.51 | 16.86 |
| – Diminisher | 1.06 | 8.12 |
| – Moderator | 19.31 | 0.67 |
| – Booster | 9.00 | 2.16 |
| – Approximator | 7.41 | 4.26 |
| – Maximizer | 0.27 | 1.40 |
| – Negation | 0.53 | 0.42 |

Table 5.2: Statistics of our spatial expression annotation in 600 random dialogues.

The basic statistics of our annotation are summarized in Table 5.2. Note that there are relatively few spatial attributes annotated, since most of them appeared inside the markables (hence not detected manually). However, a large number of spatial relations with non-obvious structures were identified.

In both expressions, we found over 1% of the subjects and 14% of the objects to be present only in previous utterances, which indicates that argument level ellipses are common and need to be resolved in visual dialogues. For spatial relations, about 30% did not have any explicit objects.

Our annotation also verified that a large portion of the spatial expressions (37% for spatial attributes and 17% for relations) accompanied modifiers. In addition, we can see that *neutrality* is the most common type of modification for spatial attributes (as in *medium gray*, *medium sized*), and *subtlety* and *uncertainty* to be the most common types for spatial relations. It is interesting to note that the frequencies of modification types vary significantly with spatial attributes and relations, except for *negation*.

Finally, less than 1% of spatial expressions were *unannotatable* based on our schema, which verifies its broad coverage. Overall, our annotation can capture important linguistic structures of visually grounded dialogues, and it is straightforward to conduct even further analyses (e.g. by focusing on specific canonical relations or modifications).

## 5.4 Model Refinement

### 5.4.1 Evaluation

In our experiments, we focus on the *reference resolution task* in OneCommon Corpus. As we discussed in Chapter 4, this can be used for probing the model's understanding of the intermediate dialogue process. As illustrated in Figure 5.1 (left), the goal of the task is to predict the referents of each markable based on the *speaker*'s perspective. To collect model predictions for all dialogues, we split the whole dataset into 10 equal-sized bins and use each bin as the test set in the 10 rounds of the experiments.

### 5.4.2 Model Architecture

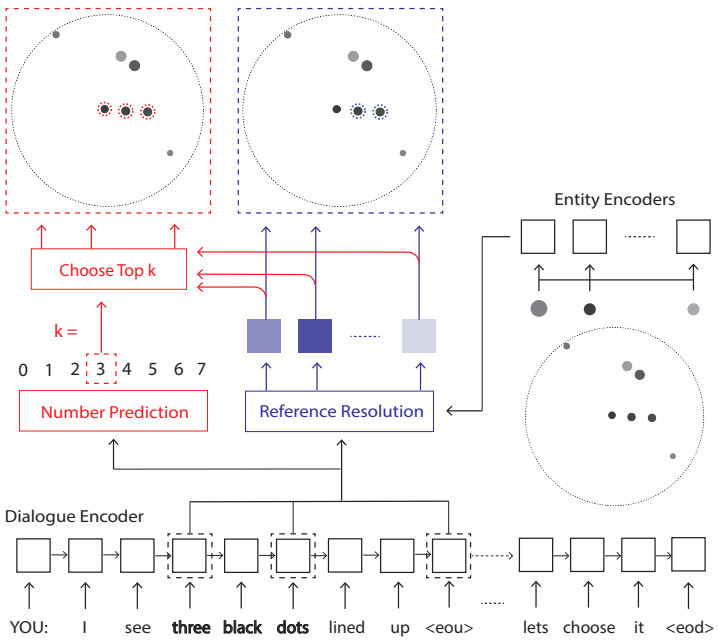



Figure 5.5: Our model architecture. REF prediction flow is shown in blue and our NUMREF prediction flow in red.

As a baseline, we use the *REF* model from Chapter 4. As shown in Figure 5.5, this model has two encoders: a *dialogue encoder* based on a simple GRU (Cho et al., 2014) and a shared *entity encoder* which outputs entity-level representation of the observation based on MLP and relational network (Santoro et al., 2017). To predict the referents, REF takes the GRU's start position of the markable, end position of the markable and end position of the utterance to compute entity-level scores and judge whether each entity is a referent based on logistic regression.

However, since the predictions are made independently for each entity, this model often predicts the wrong number of referents, leading to low performance in terms of exact match rate. To address this issue, we trained a separate module to track the *number* of referents in each markable. We formulate this as a simple classification task (between 0, 1, ..., 7) which can be predicted reliably with an average accuracy of 92%. Based on this module's prediction $k$, we simply take the top $k$ entities with the highest scores as the referents. We refer to this numerically constrained model as NUMREF.

Furthermore, we conduct feature level ablations to study the importance of each feature: for instance, we remove the xy-values from the structured input to ablate the *location* feature.

### 5.4.3 Results

| | Entity-Level Accuracy (%) | Markable-Level Exact Match (%) |
|---|---|---|
| REF | 85.71±0.23 | 33.15±1.00 |
| − location | 84.28±0.27 | 30.53±0.84 |
| − color | 83.08±0.32 | 17.09±1.04 |
| − size | 83.50±0.22 | 19.41±0.98 |
| NUMREF | **86.03±0.33** | **54.94±0.76** |
| − location | 83.35±0.26 | 49.77±0.64 |
| − color | 81.19±0.41 | 39.74±1.31 |
| − size | 82.39±0.20 | 43.40±0.67 |
| Human | 96.26 | 86.90 |

Table 5.3: Results for the reference resolution task.

We report the mean and standard deviation of the entity-level accuracy and markable-level exact match rate in Table 5.3. Compared to REF, our NUMREF model slightly improves the entity-level accuracy and significantly outperforms it in terms of exact match rate, which validates our motivation. From the ablation studies, we can see that all features contribute to the overall performance, but color and size seem to have the largest impact.

However, it remains difficult to see how and where these models struggle based on mere accuracy. For further investigation, we need more sophisticated *behavioral testing* (black-box testing) to verify whether each model has the capability of recognizing certain concepts or linguistic structures (Ribeiro et al., 2020).

## 5.5 Model Analysis

To study the current model's strengths and weaknesses in detail, we investigate whether their predictions are *consistent* with the central spatial expressions.

### 5.5.1 Spatial Attributes

First, we analyze whether the model predictions are consistent with the entity-level spatial attributes. Since most of them were confirmed to appear inside the markables (Section 5.3), we automatically detect the expressions of *color* in the markables, plot the distributions of the actual referent color, and compare the results between gold human annotation and model predictions (Figure 5.6).

From the figure, we can verify that the two distributions look almost identical for the common color expressions, and our NUMREF model seems to capture important characteristics of pragmatic expressions (same expression being used for wide range of colors) and modifications such as neutrality (*medium*) and extremity (*very dark*, *very light*).[4] Note that we observed very similar results with the *size* distributions as well.

Based on these results, we argue that the current model can capture entity-level attributes very well, including basic modification.

### 5.5.2 Spatial Relations

Next, we investigate whether the model predictions are consistent with the central spatial relations. Based on our annotation (Section 5.2), we conduct simple tests to check

[4]Spatial attributes with diminishers (such as *slightly dark*) were relatively rare and omitted in the figure.

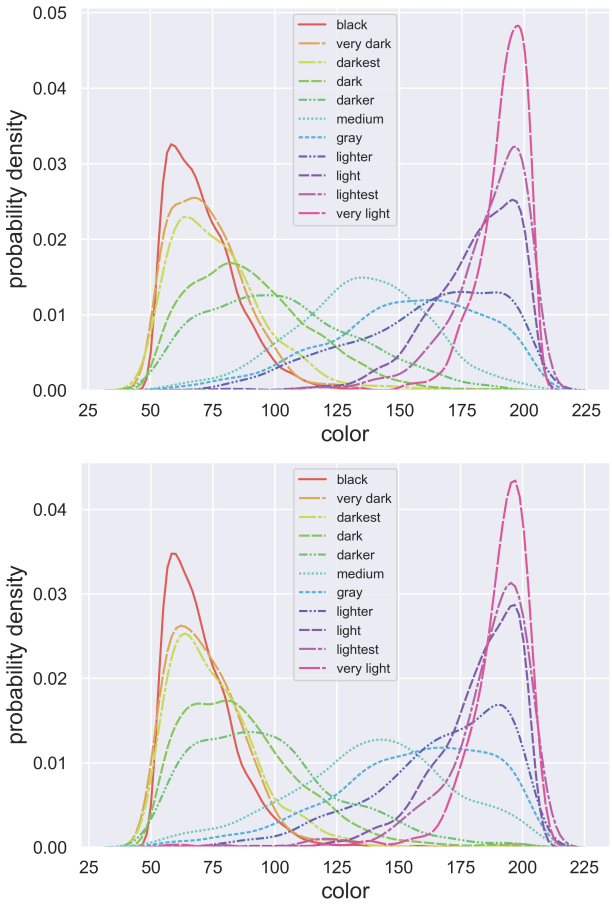



Figure 5.6: Referent color distributions. Top is human, bottom is NUMREF (smaller is darker in color axis).

whether the predicted referents satisfy each canonical relation. To be specific, our tests check for two conditions: whether the predictions are *valid* (satisfy the minimal requirements, e.g. at least 2 referents predicted for *near* relation), and if they are valid, whether the predictions actually *satisfy* the canonical relation (e.g. referents are closer than a certain threshold).

Algorithm 1 shows our test for the canonical *left* relation. Note that if no objects were annotated, we simply test whether the subject referents are on the left side of the player's view ($mean(\mathcal{S}.x) < 0$).

**Algorithm 1:** Test for *left* relation

**Input:** subject referents $\mathcal{S}$, object referents $\mathcal{O}$, boolean $no\_object$
**Output:** boolean $satisfy$, boolean $valid$
**if** $no\_object$ **then**
  $valid \leftarrow |\mathcal{S}| > 0$
  $satisfy \leftarrow valid \wedge mean(\mathcal{S}.x) < 0$
**else**
  $valid \leftarrow |\mathcal{S}| > 0 \wedge |\mathcal{O}| > 0$
  $satisfy \leftarrow valid \wedge mean(\mathcal{S}.x) < mean(\mathcal{O}.x)$
**return** $satisfy, valid$

The results of our tests are summarized in Table 5.4. We also compare with the feature ablated models to estimate the test cases which can be satisfied *without* using the corresponding features, i.e. location for *direction/proximity/region* categories, color for

| | Models | | REF | | REF-abl | | NUMREF | | NUMREF-abl | | Human | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Category | Relation | # Cases | satisfy | valid | satisfy | valid | satisfy | valid | satisfy | valid | satisfy | valid |
| Direction | *left* | 412 | 23.5 | 32.3 | 21.1 | 28.9 | **67.0** | **99.5** | 62.4 | **99.5** | 95.9 | 97.6 |
| | *right* | 468 | 28.0 | 35.5 | 24.6 | 30.8 | 67.3 | **98.7** | **68.2** | **98.7** | 95.3 | 96.4 |
| | *above* | 514 | 28.6 | 37.4 | 24.7 | 33.1 | 65.2 | 99.2 | **66.5** | **99.4** | 96.7 | 98.6 |
| | *below* | 444 | 25.2 | 34.5 | 21.6 | 27.9 | **66.0** | **99.1** | 62.2 | **99.1** | 96.4 | 96.8 |
| | *horizontal* | 37 | 54.1 | 70.3 | 27.0 | 59.5 | **59.5** | **100.0** | 51.4 | 97.3 | 91.9 | 100.0 |
| | *vertical* | 46 | 37.0 | 73.9 | 23.9 | 54.3 | 43.5 | **95.7** | **45.7** | **95.7** | 82.6 | 100.0 |
| | *diagonal* | 50 | 48.0 | 74.0 | 30.0 | 50.0 | **60.0** | **98.0** | **60.0** | **98.0** | 90.0 | 100.0 |
| | All | 1,971 | 27.8 | 37.6 | 23.4 | 31.9 | **65.5** | **99.0** | 64.1 | **99.0** | 95.5 | 97.6 |
| Proximity | *near* | 271 | 49.4 | 61.3 | 29.9 | 49.1 | **77.1** | 94.5 | 56.1 | **95.2** | 95.2 | 96.7 |
| | *far* | 27 | 29.6 | 40.7 | 33.3 | 40.7 | 77.8 | **100.0** | **92.6** | **100.0** | 96.3 | 96.3 |
| | *alone* | 111 | 36.9 | 44.1 | 45.0 | 54.1 | **68.5** | **94.6** | 67.6 | **94.6** | 91.9 | 94.6 |
| | All | 409 | 44.7 | 55.3 | 34.2 | 49.9 | **74.8** | 94.9 | 61.6 | **95.4** | 94.4 | 96.1 |
| Region | *interior* | 135 | 38.5 | 52.6 | 27.4 | 39.3 | **62.2** | 93.3 | 58.5 | **94.1** | 96.3 | 100.0 |
| | *exterior* | 62 | 40.3 | 48.4 | 40.3 | 53.2 | 80.6 | **98.4** | **87.1** | **98.4** | 98.4 | 98.4 |
| | All | 197 | 39.1 | 51.3 | 31.5 | 43.7 | **68.0** | 94.9 | 67.5 | **95.4** | 97.0 | 99.5 |
| Color | *lighter* | 147 | 23.1 | 25.9 | 6.8 | 8.2 | **84.4** | **100.0** | 57.1 | 99.3 | 97.3 | 98.0 |
| | *lightest* | 42 | 45.2 | 66.7 | 14.3 | 33.3 | **61.9** | **100.0** | 31.0 | **100.0** | 83.3 | 100.0 |
| | *darker* | 171 | 24.0 | 26.3 | 7.0 | 10.5 | **83.0** | **99.4** | 53.2 | **99.4** | 95.9 | 98.8 |
| | *darkest* | 48 | 56.2 | 64.6 | 14.6 | 33.3 | **66.7** | **100.0** | 35.4 | **100.0** | 89.6 | 97.9 |
| | *same* | 50 | 12.0 | 30.0 | 8.0 | 30.0 | **40.0** | **88.0** | 32.0 | 86.0 | 92.0 | 96.0 |
| | *different* | 14 | 64.3 | 71.4 | 71.4 | 71.4 | 64.3 | **100.0** | **78.6** | 92.9 | 92.9 | 100.0 |
| | All | 472 | 28.8 | 35.4 | 10.4 | 18.0 | **74.8** | **98.5** | 49.2 | 97.9 | 94.1 | 98.3 |
| Size | *smaller* | 213 | 27.7 | 31.5 | 7.5 | 9.9 | **80.8** | **100.0** | 59.6 | **100.0** | 98.6 | 99.5 |
| | *smallest* | 52 | 71.2 | 73.1 | 21.2 | 34.6 | **86.5** | **98.1** | 48.1 | **98.1** | 92.3 | 98.1 |
| | *larger* | 238 | 23.1 | 28.6 | 9.7 | 16.0 | **73.5** | **99.6** | 48.7 | **99.6** | 98.3 | 98.3 |
| | *largest* | 61 | 52.5 | 60.7 | 11.5 | 24.6 | **73.8** | **100.0** | 39.3 | **100.0** | 96.7 | 100.0 |
| | *same* | 103 | 34.0 | 42.7 | 18.4 | 27.2 | **80.6** | 88.3 | 65.0 | **91.3** | 98.1 | 100.0 |
| | *different* | 12 | 75.0 | 75.0 | 66.7 | 66.7 | **91.7** | **91.7** | 83.3 | 83.3 | 91.7 | 91.7 |
| | All | 679 | 33.4 | 38.7 | 12.4 | 18.9 | **78.2** | 97.8 | 54.3 | **98.1** | 97.6 | 99.0 |

Table 5.4: Canonical relation test results. We compute the *satisfy* and *valid* rate of the predictions for each canonical relation. Best scores of the models are in bold (-abl shows the corresponding feature ablated results).

*color comparison*, and size for *size comparison*.

First, we can verify that human annotation passes most of our tests, which is an important evidence of the reliability of our annotations and behavioral tests. We also confirmed that REF models often make *invalid* predictions with overall poor performance, which is consistent with our expectation.

In *direction*, *proximity* and *region* categories, we found that NUMREF model performs on par or only marginally better than its ablated version (and even underperforms it for simple relations like *right* and *above*): these results indicate that current model is still incapable of leveraging locational features to make more consistent predictions.[5]

In *color/size comparison*, NUMREF performs reasonably well, outperforming all other models: this indicates that the model can not only capture but also *compare* entity-level attributes to a certain extent. However, there is still room left for improvement in almost all relations. It is also worth noting that *size comparison* may be easier, as the range of size is limited (only *6* compared to *150* for color).

Overall, we conclude that current models still struggle in capturing most of the inter-entity relations, especially those related to placements.

[5]For relations like *far* and *different color*, ablated model may be better simply because referents tend to be more distant/dissimilar when predictions are closer to random.

### 5.5.3 Further Analyses

| Linguistic Factors | # Cases | NUMREF | Human |
|---|---|---|---|
| Strong modification | 149 | 76.51 | 95.97 |
| Neutral | 3,094 | 70.46 | 95.77 |
| Weak modification | 490 | 66.12 | 95.10 |
| Inter-utterance subject | 14 | 57.14 | 92.86 |
| Inter-utterance object | 265 | 72.08 | 94.72 |
| No object | 1,127 | 74.45 | 92.99 |
| Ignorable object | 1,805 | 69.64 | 97.23 |
| Unignorable object | 796 | 65.33 | 96.11 |
| All | 3,728 | 70.17 | 95.71 |

Table 5.5: Satisfy rates classified by linguistic factors.

Finally, we conduct further analyses to study other linguistic factors that affect model performance. Table 5.5 shows the results of our relation tests classified by notable linguistic structures.

In terms of modification, we can confirm that human performance is consistently high, while the model performs best for strong modification (modified by *boosters* or *maximizers*), decently for neutrals (*moderators* or no modification), and worst on weak modification (*diminishers* or *approximators*). This indicates that large, conspicuous features are easier for the model to capture compared to small or more ambiguous features.

In terms of subject/object properties, human performance is also consistently high. In contrast, model performance is significantly worse for subject ellipsis (*inter-utterance subject*), while remaining high for object ellipsis and *no object* cases.

We also hypothesize that a large portion of the relations can actually be satisfied *without* considering the objects, e.g. by simply predicting very dark dots as the subjects when the relation is *darker* or *darkest*. To distinguish such easy cases, we consider a relation as *ignorable object* if the relation can be satisfied even if we ignore the objects (i.e. remove all object relations) based on gold referents. Our result verifies that there are indeed many cases of *ignorable object*, and they seem slightly easier for the model to satisfy.

| Models | | NUMREF | | Human | |
|---|---|---|---|---|---|
| value | mod-type | diff. | # valid | diff. | # valid |
| xy-value | strong | 86.06 | 39 | 89.15 | 37 |
| | neutral | 80.92 | 1,586 | 73.52 | 1,558 |
| | weak | 80.35 | 200 | 53.53 | 198 |
| color | strong | 66.23 | 15 | 91.80 | 15 |
| | neutral | 56.98 | 234 | 60.14 | 232 |
| | weak | 37.73 | 68 | 28.55 | 66 |
| size | strong | 3.60 | 8 | 4.29 | 8 |
| | neutral | 2.67 | 337 | 2.70 | 320 |
| | weak | 1.95 | 105 | 1.58 | 104 |

Table 5.6: Absolute differences of feature values in comparative relations (number of valid predictions shown in shade).

In Table 5.6, we study the effect of modification based on the *absolute difference*

between subject and object features in comparative relations.[6]

In human annotation, the absolute difference naturally increases as the modification gets stronger. While model predictions also show this tendency, their results seem less sensitive to modification (particularly for locational features, i.e. xy-value) and may not be reflecting their full effect.

## 5.6 Related Work

Linguistic structure plays a critical role in dialogue research. From theoretical aspects, various dialogue structures have been studied, including discourse structure (Stent, 2000; Asher et al., 2003), speech act (Austin, 1962; Searle, 1969) and common grounding (Clark, 1996; Lascarides and Asher, 2009). In dialogue system engineering, various linguistic structures have been considered and applied, including syntactic dependency (Davidson et al., 2019), predicate-argument structure (Yoshino et al., 2011), ellipsis (Quan et al., 2019; Hansen and Søgaard, 2020), intent recognition (Silva et al., 2011; Shi et al., 2016), semantic representation/parsing (Mesnil et al., 2013; Gupta et al., 2018) and frame-based dialogue state tracking (Williams et al., 2016; El Asri et al., 2017). However, most prior work focus on dialogues where information is not grounded in external, perceptual modality such as vision. In this work, we propose an effective method of analyzing linguistic structures in visually grounded dialogues.

Recent years have witnessed an increasing attention in visually grounded dialogues (Zarrieß et al., 2016; de Vries et al., 2018; Alamri et al., 2019; Narayan-Chen et al., 2019). Despite the impressive progress on benchmark scores and model architectures (Das et al., 2017b; Wu et al., 2018; Kottur et al., 2018; Gan et al., 2019; Shukla et al., 2019; Niu et al., 2019; Zheng et al., 2019; Kang et al., 2019; Murahari et al., 2019; Pang and Wang, 2020), there have also been critical problems pointed out in terms of dataset biases (Goyal et al., 2017; Chattopadhyay et al., 2017; Massiceti et al., 2018; Chen et al., 2018; Kottur et al., 2019; Kim et al., 2020; Agarwal et al., 2020) which obscure such contributions. For instance, Cirik et al. (2018) points out that existing dataset of reference resolution may be largely solvable *without* recognizing the full referring expressions (e.g. based on object categories only). To circumvent these issues, we focused on our OneCommon Corpus where the visual contents are simple (exploitable categories are removed) and well-balanced (by sampling from uniform distributions) to minimize dataset biases.

Although various probing methods have been proposed for models and datasets in NLP (Belinkov and Glass, 2019; Geva et al., 2019; Kaushik et al., 2020; Gardner et al., 2020; Ribeiro et al., 2020), fine-grained analyses of visually grounded dialogues have been relatively limited. Instead, Kottur et al. (2019) proposed a diagnostic dataset to investigate model's language understanding: however, their dialogues are generated artificially and may not reflect the true nature of visual dialogues. Shekhar et al. (2019) also acknowledges the importance of linguistic analysis but only dealt with coarse-level features that can be computed automatically (such as dialogue topic and diversity). Most similar and related to our research are Yu et al. (2019a) and Udagawa and Aizawa (2020), where they conducted additional annotation of reference resolution in visual dialogues: however, they still do not capture more sophisticated linguistic structures such as predicate-argument structure, modification and ellipsis.

Finally, spatial language and cognition have a long history of research (Talmy, 1983; Herskovits, 1987). In computational linguistics, Kordjamshidi et al. (2010) and Pustejovsky et al. (2015) developed the task of spatial role labeling to capture spatial information in text: however, they do not fully address the problem of annotation reliability nor

[6] *Left/right* for x-value, *above/below* for y-value, *lighter/darker* for color and *smaller/larger* for size.

grounding in external visual modality. In computer vision, the VisualGenome dataset (Krishna et al., 2017b) provides rich annotation of spatial scene graphs constructed from raw images, but not from raw dialogues. Ramisa et al. (2015) and Platonov and Schubert (2018) also worked on modelling spatial prepositions in single sentences. To the best of our knowledge, our work is the first to apply, model and analyze spatial expressions in visually grounded dialogues at full scale.

## 5.7 Discussion and Conclusion

In this study, we focused on the (annotated) OneCommon Corpus as a suitable testbed for fine-grained language understanding in visually grounded dialogues. To analyze its linguistic structures, we proposed a novel framework of annotating spatial expressions in visual dialogues. We showed that our annotation can be conducted reliably and efficiently by leveraging referring expressions prevalent in visual dialogues, while capturing important linguistic structures such as predicate-argument structure, modification and ellipsis. Although our current analysis is limited to this domain, we expect that upon appropriate definition of spatial expressions, argument roles and canonicalization, the general approach can be applied to a wider variety of domains: adapting and validating our approach in different domains (especially with more realistic visual contexts) are left as future work.

Secondly, we proposed a simple idea of incorporating *numerical constraints* to improve exophoric reference resolution. We expect that a similar approach of identifying and incorporating semantic constraints (e.g. coreferences and spatial constraints) is a promising direction to improve the model's performance even further.

Finally, we demonstrated the advantages of our annotation for investigating the model's understanding of visually grounded dialogues. Our tests are completely agnostic to the models and only require referent predictions made by each model. By designing simple tests like ours (Sections 5.5.1 and 5.5.2), we can diagnose the model's performance at the granularity of canonical attributes/relations under consideration: such analyses are easy to extend (by adding more tests) and critical for verifying what capabilities current models have (or do not have). Based on further analyses (Section 5.5.3), we also revealed various linguistic structures that affect model performance: we expect that capturing and studying such effects will be essential for advanced model probing in visual dialogue research.

Overall, we expect our framework and resource to be fundamental for conducting sophisticated linguistic analyses of visually grounded dialogues, involving advanced common grounding and symbol grounding.

# Chapter 6

# Task Generalization under Dynamic Context

Common grounding is the process of creating and maintaining mutual understandings, which is a critical aspect of sophisticated human communication. While various task settings have been proposed in existing literature, they mostly focus on creating common ground under static context and ignore the aspect of *maintaining* them overtime under *dynamic* context. In this chapter, we propose a novel task setting to study the ability of both creating and maintaining common ground in dynamic environments. Based on our minimal task formulation, we collected a large-scale dataset of 5,617 dialogues to enable fine-grained evaluation and analysis of various dialogue systems. Through our dataset analyses, we highlight novel challenges introduced in our setting, such as the usage of complex *spatio-temporal expressions* to create and maintain common ground. Finally, we conduct extensive experiments to assess the capabilities of our baseline dialogue system and discuss future prospects of our research.

## 6.1 Introduction

Common grounding is the process of creating, repairing and updating mutual understandings (i.e. *common ground*), which is a critical aspect of sophisticated human communication (Clark, 1996). Humans can *create* substantial common ground by expressing various information in natural language, which can be clarified or *repaired* to resolve misunderstandings at essential levels of detail. Furthermore, as the situation changes and relevant information gets outdated, humans can *update* their common ground accordingly by discarding old information and acquiring new ones. Such ability plays a vital role in sustaining collaborative relationships and adapting to emerging problems in nonstationary, real-world environments.

However, despite the wide variety of tasks proposed in existing literature (Fang et al., 2015; Zarrieß et al., 2016; De Vries et al., 2017; Udagawa and Aizawa, 2019; Haber et al., 2019), they mostly focus on creating common ground under *static* (time-invariant) context and ignore their *dynamic* aspects. While some recent dialogue tasks deal with dynamic information, they often lack suitable evaluation metrics (Pasunuru and Bansal, 2018), *context updates* in the course of the dialogue (Alamri et al., 2019) or diverse dynamics of the environment itself (de Vries et al., 2018; Suhr et al., 2019; Narayan-Chen et al., 2019; Thomason et al., 2019; Moon et al., 2020). Therefore, it remains unclear how well existing dialogue systems can adapt to the diversely changing situations through advanced common grounding.

To address this problem, we propose a novel dialogue task based on three design choices (Section 6.2):

First, we formulate a novel *sequential collaborative reference task* as a temporal generalization of the collaborative reference task proposed in Chapter 3. In our formulation, the goal of the agents is generalized to track and select the common entity *at multiple*

*timesteps*, while the agents' observations change dynamically between each timestep. This setting requires both *creation* and *maintenance* of common ground, whilst enabling clear evaluation based on the length of successful timesteps.

Secondly, we focus on synthesizing the *entity movements*, as popularized in the recent video understanding benchmarks (Girdhar and Ramanan, 2020; Yi et al., 2020; Bakhtin et al., 2019). By leveraging such synthetic dynamics, we can minimize undesirable biases, maximize diversity and enable fully controlled evaluation and analysis.

Finally, we build upon our OneCommon Corpus to introduce natural difficulty of common grounding with minimal task complexity. To be specific, we represent entity attributes and their temporal dynamics based on *continuous* real values to introduce high ambiguity and uncertainty. In addition, we consider a *partially-observable* setting where each agent only has a partial view of the environment, which introduces various misunderstandings and partial understandings that need to be resolved.

Based on this task design, we collected a large-scale dataset of 5,617 dialogues (including over 65K utterances) through careful crowdsourcing on Amazon Mechanical Turk (Section 6.3).

We show an exemplary dialogue of our task in Figure 6.1. Since the environment is dynamic, humans rely on various *spatio-temporal expressions* to express entity states at different timesteps ("started off on the left", "ends to the right") or how they changed dynamically ("moves very quickly", "come towards the left") to create common ground. Furthermore, in later turns, humans often leverage their *previous common ground* ("still see the same one?", "crosses underneath our old one") to update their common ground more reliably and efficiently. We conduct detailed analyses of the dataset to study such strategies in Section 6.4.

In our experiments (Section 6.5), we train a neural-based dialogue system based on Udagawa and Aizawa (2020). Through our extensive evaluation and analysis, we assess the current model's strengths as well as important limitations and demonstrate huge room left for further improvement.

Overall, the contributions of this chapter are as follows:

- We proposed a novel dialogue task to study common grounding in dynamic environments.
- We collected a large-scale dataset of 5,617 dialogues to develop and test various data-driven models.
- Our detailed dataset analyses highlight novel challenges introduced in our setting.
- We conduct extensive evaluation and analysis of a simple yet strong baseline dialogue system.

## 6.2 Task Formulation

In this section, we review the collaborative reference task from OneCommon Corpus (OCC in short) and formulate our *sequential* counterpart as its temporal generalization.

### 6.2.1 Collaborative Reference Task

Based on Udagawa and Aizawa (2019), a *collaborative reference task* is a multi-agent cooperative game with entities $E = \{e_1, e_2, ..., e_m\}$ and agents $A = \{a_1, a_2, ..., a_n\}$. Each agent $a_j \in A$ has an observation of entities $obs_j(E)$ and can exchange information with other agents in natural language. At the end of the game, each agent selects one of the observable entities, and the game is *successful* if and only if all the agents selected

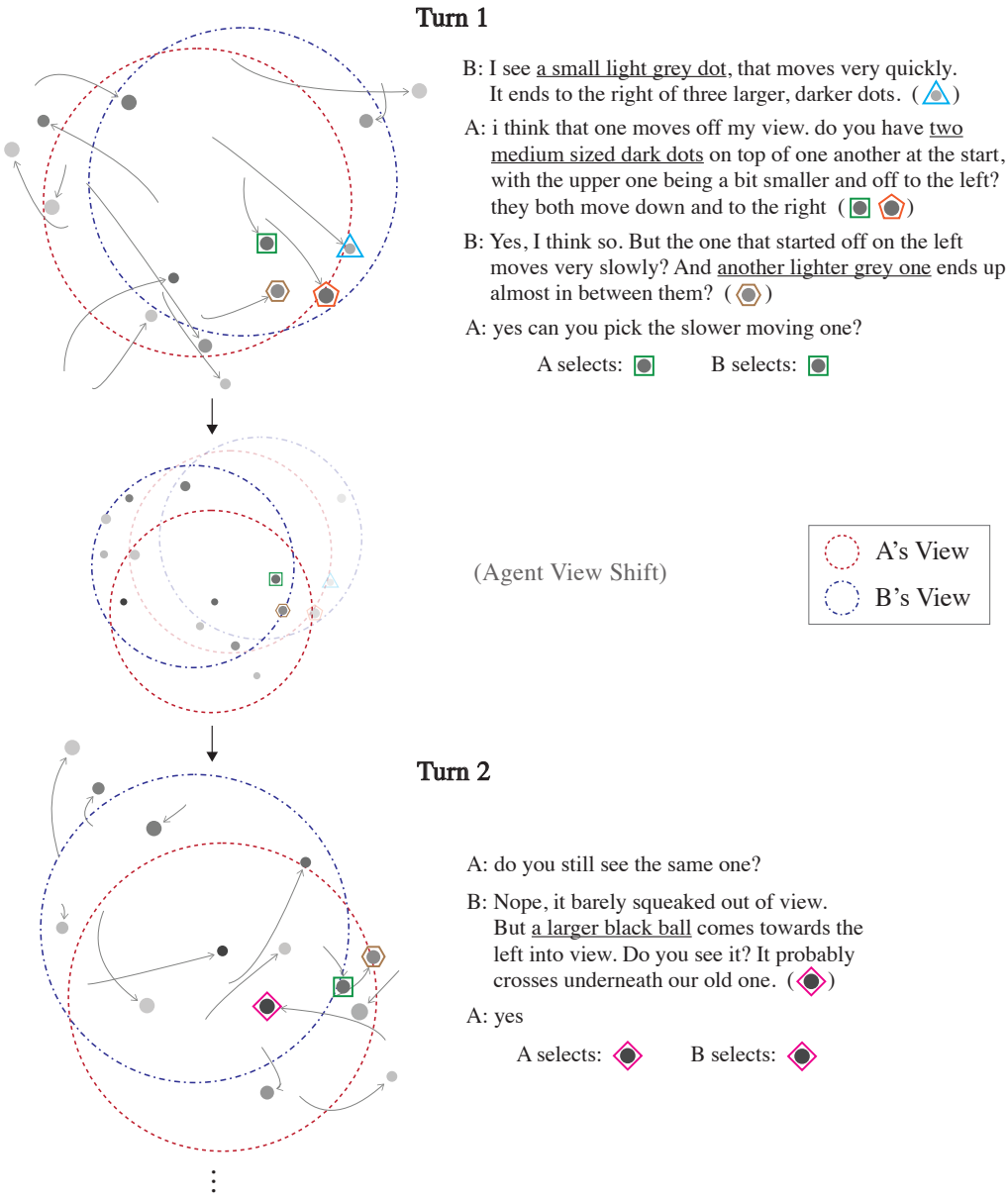



Figure 6.1: Example dialogue of our sequential collaborative reference task. Each agent has a partial view of a 2-D plane with synthetic entities (grayscale dots of various sizes). *During* each turn, the entities move randomly on the 2-D plane. *At the end* of each turn, the agents communicate with each other to find and select one of the same, common entities. *After* each turn (if the selections match), both agents' views shift randomly and the next turn begins.

the same entity. This can be considered as a general framework for evaluating accurate *mutual recognition* of a common entity, which is often a critical step in general common grounding.

One main feature of OCC is that they represented all entity attributes (color, size and location on a 2-D plane) based on *continuous* real values. Unlike discrete/categorical attributes, this introduces high ambiguity and uncertainty to be expressed in symbolic natural language. In addition, they introduced *partial-observability* where each agent only has a partial view of the 2-D plane, which requires collaborative resolution of various misunderstandings.

However, this current task formulation assumes each observation to be *static* and can

only evaluate the ability of *creating* common ground.

### 6.2.2 Sequential Collaborative Reference Task

To address this limitation, we generalize each observation to be *dynamic* and collaborative reference to be *sequential*. Specifically, each agent $a_j \in A$ now receives observation $obs_j(E, t)$ at each timestep $t \in [t_0, \infty)$, and the agents' goal is to communicate in natural language to select the same entity at multiple timesteps $t_1, t_2, ... \in (t_0, \infty)$.[1] At each selection timestep $t_k$ ($k \in \mathbb{N}$), $a_j$ must select one entity observable at $t_k$ but has all previous observations up to $t_k$, $\{obs_j(E, t) | t \in [t_0, t_k]\}$. The game ends when the selections no longer match at timestep $t_{k'}$ ($k' \in \mathbb{N}$): therefore, the success at $t_1$ measures the ability of *creating* common ground, and the length of successful timesteps (LST) $k' - 1$ measures the ability of *maintaining* them. This is a general framework for evaluating both creation and maintenance of mutual entity recognition in dynamic environments.

Based on this task formulation, we propose a minimal task setting extending OCC and incorporate dynamic change of the entity *locations*.

We refer to each time range $[t_{k-1}, t_k]$ as *turn* $k$. During each turn, we change the location of each entity $e_i \in E$ based on a simple parameterized movement, where the *trajectory* is determined by a quadratic Bézier curve (Bézier, 1974).[2] See Figure 6.2 for an illustration, where $r_1$, $r_2$ are parameters of *distance* and $\theta_{k-1}$, $\Delta\theta$ represent *angles*. We sample $r_1$, $r_2$, $\Delta\theta$ from fixed uniform distributions each turn and update $\theta_k$ as $\theta_k \leftarrow \theta_{k-1} + \Delta\theta$ ($\theta_0$ is initialized randomly). This way, we can generate diverse, unbiased, coherent and fully controllable dynamics of the environment.

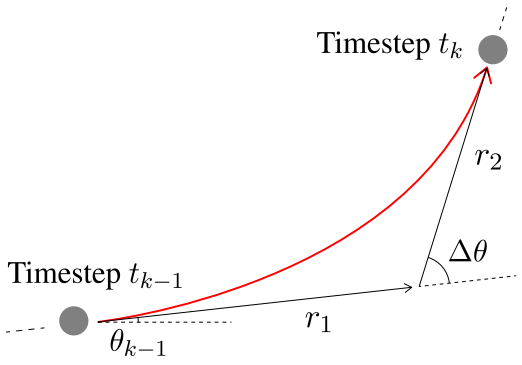



Figure 6.2: Illustrated movement of each entity in turn $k$.

To enable fair comparison with OCC, we limit the number of agents to 2 and set the circular agent views to have the same diameter as OCC. At each selection timestep $t_k$, we ensure that each agent has 7 observable entities with only 4, 5 or 6 of them in common, which is also identical to OCC. Finally, we sample all entity attributes (color, size and initial location) from the same uniform distributions as OCC with minimal modifications.[3] Therefore, we expect the (distribution of) observations at $t_k$ to be similar and enable mostly fair comparison with OCC (Sections 6.4 and 6.5).

To ensure task difficulty, we also shift the *perspective* of each agent after each successful turn (see Figure 6.1) so that the overlapping regions differ every turn. The same dot is prohibited from staying in common for over 3 consecutive selection timesteps, requiring frequent updates of common ground. Finally, we limit the maximum number of turns to 5 for practical purposes (hence the maximum LST is 5 in each game).

[1] We assume $t_{k-1} < t_k$ for all $k \in \mathbb{N}$.

[2] Its *speed* is proportional to the length of the trajectory.

[3] To be specific, we set the minimum distance between entities (at $t_k$) and the possible range of entity size to be slightly different to avoid entity overlapping during movements.

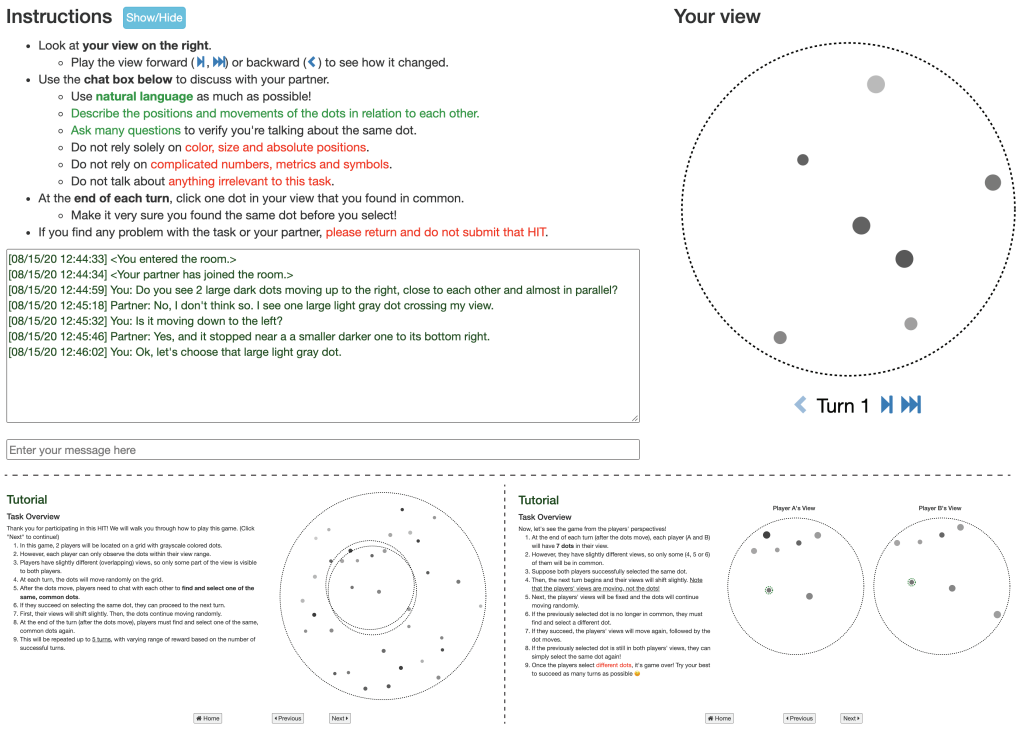



Figure 6.3: (Top) Our dialogue interface. During the game, animations up to the current turn could be replayed anytime using the forward/backward buttons. (Bottom) Sample screenshots from our tutorial on the *task setting*.

## 6.3 Dataset Collection

To collect large-scale, high-quality dialogues, we conducted careful crowdsourcing on Amazon Mechanical Turk. The web application is based on the CoCoA framework (He et al., 2017), and we used Scalable Vector Graphics (SVG) to animate entity movements and parallel shifts of the agent perspectives. Before working on our task, crowd workers were required to take a brief tutorial on the task setting, dialogue interface and instructions. Sample screenshots of our dialogue interface and tutorial are shown in Figure 6.3: note that animations up to the current turn could be replayed anytime for the ease of playing the game.[4]

To ensure worker quality, we required crowd workers to have more than 500 completed HITs and acceptance rates higher than 99%. To encourage success, we rewarded \$0.25 for every successful turn plus additional bonuses for longer LST achieved (up to \$0.25 if LST $= 5$). Finally, we manually reviewed all submitted works and excluded dialogues which clearly violated the instructions (e.g. relying on premature guessing or other ineffective strategies[5]). We did not exclude dialogues based on task failures (even if LST $= 0$), as long as they were based on valid strategies.

To solicit linguistic/strategic variety, we generally used a unique environment for each game. However, if the task was unsuccessful (i.e. LST $= 0$), we allowed the environment to be reused in another game. This way, we can expect to eventually collect successful (LST $> 0$) dialogues for the relatively difficult environments as well.

Overall, we collected 5,804 dialogues, and after the reviewing process, we were

[4]This also allows us to ignore the disadvantage of imperfect human memories in comparison to machines.

[5]Typical examples include strategies relying solely on color, size and absolute positions in the agent's view.

left with 5,617 qualified dialogues. We refer to this dataset as **Dynamic-OneCommon Corpus (D-OCC)**. Note that our dataset is currently in English, but the dataset collection procedure is language-agnostic and can be applied in any other languages.

## 6.4 Dataset Analysis

Next, we conduct detailed analyses of the dataset to study human common grounding strategies under dynamic context. Whenever possible, we give comparative analyses with OCC to highlight the effect of dynamic factors introduced in D-OCC.

### 6.4.1 Overall Statistics

| Statistics | OCC | D-OCC |
|---|---|---|
| Total dialogues | 6,760 | 5,617 |
| Uttrances per dialogue | 4.8 | 11.7 |
| Tokens per utterance | 12.4 | 10.3 |
| Duration per dialogue (minutes) | 2.1 | 5.7 |
| Unique workers | N/A | 462 |
| Avg. LST | - | 3.31 |
| Avg. completed turns | - | 3.77 |
| Unique tokens | 3,621 | 3,895 |
| Occupancy of rare tokens (%) | 1.4 | 1.0 |
| Overlap of all tokens (%) | 29.4 | |
| Overlap w/o rare tokens (%) | 53.0 | |

Table 6.1: Statistics of OCC and D-OCC datasets.

First, we summarize the overall statistics of OCC and D-OCC in Table 6.1.

In total, OCC and D-OCC have a comparable number of dialogues. However, dialogues can be much longer in D-OCC, since collaborative reference is repeated multiple times. On average, utterance lengths are slightly shorter in D-OCC: this can be attributed to the increased (relative) frequency of short utterances like acknowledgements and *shortened* subsequent responses (e.g. "same again?" = "select the same black dot again?").[6] Note that long, complex utterances are also common in our dataset, as seen in Figure 6.1. Overall, we found 462 unique workers participated in D-OCC, which indicates reasonable diversity at the *player* level as well.

In terms of LST, the overall average was 3.31 with over half (53.5%) of the dialogues succeeding all 5 turns. This suggests that humans can solve the task reliably through sophisticated common grounding. After filtering dialogues with poor/careless workers (whose avg. LST $< 2$), we observed a slight improvement up to 3.57. If we only focus on the top 10 workers (with at least 10 tasks completed), avg. LST was significantly higher reaching 4.24. These results indicate that (at least potentially) much higher human ceiling performance can be achieved. Note that if we include the unsuccessful turns in 46.5% of the dialogues, the average of all completed turns was slightly longer (3.77) in our dataset.

Finally, we found that both datasets have a relatively small vocabulary size as well as the occupancy of *rare tokens* (used less than 10 times in the dataset).[7] This indicates minimal complexity at the *lexical* level, as observed in Udagawa and Aizawa (2019).

[6] In fact, utterances with less than 5 tokens were almost twice more frequent in D-OCC (33.8%) than OCC (17.6%).

[7] Occupancy is computed based on the proportion of total frequencies (TF), i.e. *TF of rare tokens* / *TF of all tokens*.

We also found that the two datasets have a large vocabulary overlap, which is expected as D-OCC extends the setting of OCC.

| Reference | Examples | Frequency / Cohen's $\kappa$ |
|---|---|---|
| Current State | It's to the right of where the grey one ended up for me after moving up and left.<br>Now I have another triangle / Does it land next to two smaller gray dots?<br>Does it have a lighter one below and to the left when they stop?<br>Two similar shades close to each other (*implicit*) | 23.8%<br>/ 0.91 |
| State Change | a small dark one traveling southwest / 2 other dots following it<br>Do you have two dark med-size dots move slowly apart as they drift right?<br>I have a large pale grey that moves down but starts out curving to the right and then takes a sharp turn to the south east | 32.7%<br>/ 0.97 |
| Previous State | I still see the larger gray one that was next to it in the previous turn.<br>I have the smaller dot that started out below it to the left.<br>Before it moves, is there a lighter gray dot down and to the right of it? | 5.5%<br>/ 0.79 |

Table 6.2: Spatio-temporal expressions. Keywords (such as *tense*, *events* and *motion verbs*) are underlined.

| Degree Modifiers | OCC | D-OCC | # Keywords | Usage in D-OCC |
|---|---|---|---|---|
| Diminishers | 9.2 | 8.9 | 10 | **slightly** curves up |
| Moderators | 1.3 | 0.9 | 6 | **fairly** quickly |
| Boosters | 9.8 | 6.1 | 27 | **extremely** slowly |
| Approximators | 10.2 | 6.4 | 34 | **almost** collides with |
| Maximizers | 4.3 | 4.2 | 37 | **perfectly** straight |

Table 6.3: Average occurrences of degree modifiers per 100 utterances (estimated based on keywords).

### 6.4.2 Spatio-Temporal Expressions

At the utterance level, we observed an extensive usage of *spatio-temporal expressions* which are characteristic in dynamic environments. To study the frequency of such expressions, we manually annotated 100 dialogues in D-OCC with LST $\geq 2$ (focusing on the more successful strategies).

Specifically, we detect whether each utterance contains 3 types of spatio-temporal expressions:[8]

- Reference to **current state** describes location of entities at the end of the current turn (i.e. timestep $t_k$ if the utterance is in turn $k$).
- Reference to **state change** describes temporal change of entity locations (i.e. movements).
- Reference to **previous state** describes entity locations at previous timestep $t$ (where $t < t_k$).

We show examples and estimated frequencies of spatio-temporal expressions in Table 6.2. We also computed the agreement of our annotation based on 50 dialogues with 3 annotators, which we found to be reliable based on Cohen's $\kappa$ (Cohen, 1968).

[8]Note that a single utterance may contain none or multiple types of such expressions, and expressions of color, size or possession are not considered as spatio-temporal expressions.

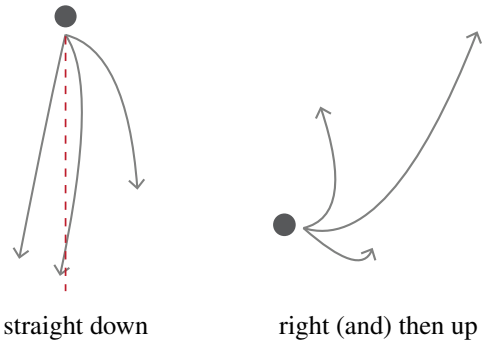



Figure 6.4: Pragmatic expressions of movements.

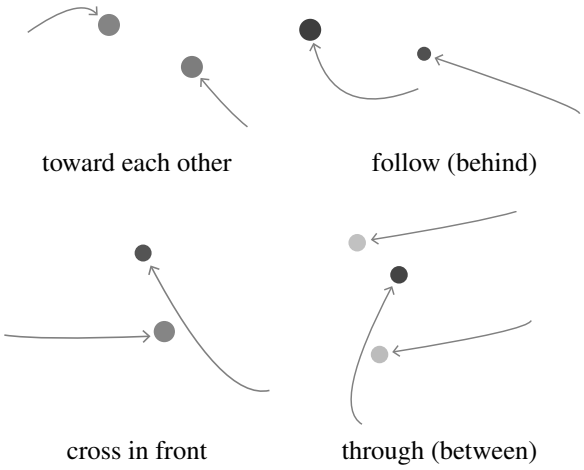



Figure 6.5: Expressions of multiple entity interactions.

Based on this result, we found that reference to *state change* is the most widely used strategy, which could be simple as "moves northwest" or more complex as in Table 6.2. Reference to *previous state* is much less frequent compared to other types but still observed in many dialogues. Note that humans distinguish *previous* and *current* states in various ways, including temporal expressions ("was", "now"), motion verbs ("started out", "landed") and implicit/default reasoning.

We also found that expressions are often *nuanced* and *pragmatic*, which are characteristic under continuous and partially-observable context (Udagawa and Aizawa, 2019). Nuances are typically expressed by the *degree modifiers* to convey subtle differences in location, movements, confidence, etc. Following Paradis (2008), we categorize them into 2 main types (and 5 subtypes): *scalar modifiers* used for concepts in a range of scale (*diminishers*, *moderators*, *boosters*) and *totality modifiers* used for concepts with definite boundaries (*approximators*, *maximizers*). See Table 6.3 for examples and the estimated occurrences of such modifiers in OCC and D-OCC.[9] Based on these results, we can verify that there are comparable numbers of various degree modifiers in D-OCC as well, which are used effectively to cope with complex ambiguity and uncertainty.

In Figure 6.4, we show examples of *pragmatic* expressions which require pragmatic (non-literal) interpretations (Monroe et al., 2017). For instance, trajectories of the expression "straight down" may not indicate vertical lines in the literal sense (e.g. could be curving or leaning to the left). Similarly, the expression of "(moving) right and (then)

[9]Following the prior analysis in OCC, we manually curated keyword-based dictionaries of such modifiers (based on unigrams and bigrams) while removing polysemous words (such as *little*, *about*, *too*, etc).

| Dataset | Turn | Previous Target | Success Rate (%) #Shared=4 | #Shared=5 | #Shared=6 | Utterances per Turn | Tokens per Utterance |
|---|---|---|---|---|---|---|---|
| OCC | 1st | - | 65.8 | 77.0 | 87.0 | 4.8 | 12.4 |
| D-OCC | 1st | - | 73.4 | 82.0 | 87.6 | 3.2 | 11.0 |
| | ≥2nd | ✓ | 95.4 | 97.0 | 97.8 | 2.3 | 5.9 |
| | | ✗ | 81.7 | 88.4 | 91.6 | 3.5 | 11.7 |

Table 6.4: Turn-level statistics of OCC and D-OCC. ✓ denotes cases where the previous target stays in common and ✗ denotes it left at least one agent's view. Note that # shared entities are 4, 5 or 6 at selection timesteps (Section 6.2.2).

| Previous Target | Examples | Frequency (%) |
|---|---|---|
| Stay (✓) | I still see the same dot / same one again?<br>I still have all three dots from the line before<br>Left my screen, but may have come back traveling left to right? | 36.8 |
| Leave (✗) | I lost the last one / both are gone for me<br>I lost the light one but still see the darker one that was on its left.<br>similar size black dot that barely moves? (*implicit*) | 63.2 |

Table 6.5: Comparison of utterances when the previous target stays in common (✓) or not (✗).

up" may be used for diverse movements ending up in various locations (e.g. even below the initial location!). While such expressions more or less deviate from literal semantics, they are pragmatically sufficient to convey the speaker's intention (i.e. identify the target among the distractors) (Grice, 1975): alternatively, the speaker may need to choose different expressions for the same movement depending on the context (distractors).

We also show exemplary expressions of multiple entity interactions in Figure 6.5, which demonstrate interesting pragmaticality as well. For instance, "toward each other" may be used for trajectories moving in *orthogonal* (rather than opposite) directions for the most of the time.

Overall, our analyses of spatio-temporal expressions reveal advanced language understanding and generation required in D-OCC, regardless of the task/lexical simplicity.

### 6.4.3 Turn-Level Strategies

Finally, we study and compare human strategies at different timesteps (in different turns). Table 6.4 shows detailed statistics of the dataset in the initial turn and later turns, where *creation* and *maintenance* of common ground are required, respectively. Note that we also distinguish later turns based on whether the previous selection (i.e. *previous target*) stays in common (✓) or leaves at least one agent's view (✗): former cases can *retain* the same common ground but the latter cases require an *update* of common ground.

First, if we focus on the 1st turn, we can verify that success rates are consistently higher in D-OCC than OCC, especially in difficult cases when the number of shared entities is smaller. This indicates that humans can create common ground more accurately by leveraging dynamic information (e.g. entity movements) unavailable in OCC.

In later turns, we found that human performance is near perfect with shorter dialogues in ✓ cases (when the previous target stays in common). This is natural because they can simply retain common ground and repeat the same selection. Notably, human performance is consistently higher than the 1st turn *even in* ✗ cases (when the previous target is no longer in common), which verifies that humans can leverage previous common ground to *update* common ground more reliably as well.

We show example utterances of ✓ and ✗ cases in Table 6.5. Note that the previous target may temporarily leave the view and come back in ✓ cases, which occasionally makes even *retainment* of the same common ground non-trivial. In ✗ cases, humans either inform about the lost entities explicitly or *implicitly*, e.g. by ignoring old entities and starting to focus on the new ones.

## 6.5 Experiments

Finally, we conduct extensive experiments to assess our baseline model's capability of common grounding in dynamic environments.

### 6.5.1 Evaluation

To study the model's capability from various aspects, we design 3 (sub)tasks based on D-OCC.

First, we evaluate the model's ability of *recognizing* common ground based on the **target selection task**, originally proposed for OCC. This is an important subtask of (sequential) collaborative reference, where the model is given one player's observation and the (ground-truth) dialogue history to predict which target was selected by the player. Since there can be multiple selections in D-OCC, the model makes predictions at the end of each turn $k$ (at timestep $t_k$). The number of entities observable at $t_k$ is fixed at 7 for both OCC and D-OCC (Section 6.2.2), so this is a simple classification task evaluated based on accuracy.

Secondly, we estimate the model's ability of *creating* and *maintaining* common ground based on the **selfplay dialogue task**, where each model plays the full sequential collaborative reference task against an identical copy of itself. While this evaluation has the advantage of being scalable and automatic, succeeding on this setting is only *necessary* for human-level common grounding and not *sufficient*, since the model may only be able to coordinate with itself (and not with real humans).

Thirdly, we conduct **human evaluation** to test the model's ability of playing sequential collaborative reference against real human workers on AMT. Due to the high cost of this evaluation, we only focus on the top 3 variants of our baseline ranked by avg. LST in the selfplay dialogue task.

### 6.5.2 Model Architecture

For a fair comparison with prior work, we implement our baseline model following the OCC models in Udagawa and Aizawa (2020). The overall model architecture is shown in Figure 6.6.

To encode the dialogue tokens throughout the turns, we use a unidirectional GRU (Cho et al., 2014). To encode the observation during turn $k$, we first split the animation of entity movements into 10 frames and the agent view shift into 5 frames. Then, we process each observation frame based on the *spatial* encoder, followed by the *temporal* encoder to integrate these outputs.

The spatial encoder is used to extract *spatial features* and *meta features* from each observation frame. Spatial features represent the spatial attributes of each entity (color, size and location in the frame), which are encoded using an MLP and a relation network (Santoro et al., 2017). The relation network is used to represent the spatial attributes relative to a subset of entities $\tilde{E} \subset E$, which could be *all entities* observable in turn $k$ ($E_{all}$) or *selectable entities* visible at $t_k$ ($E_{sel}$). Hence, the spatial features of $e_i$ are

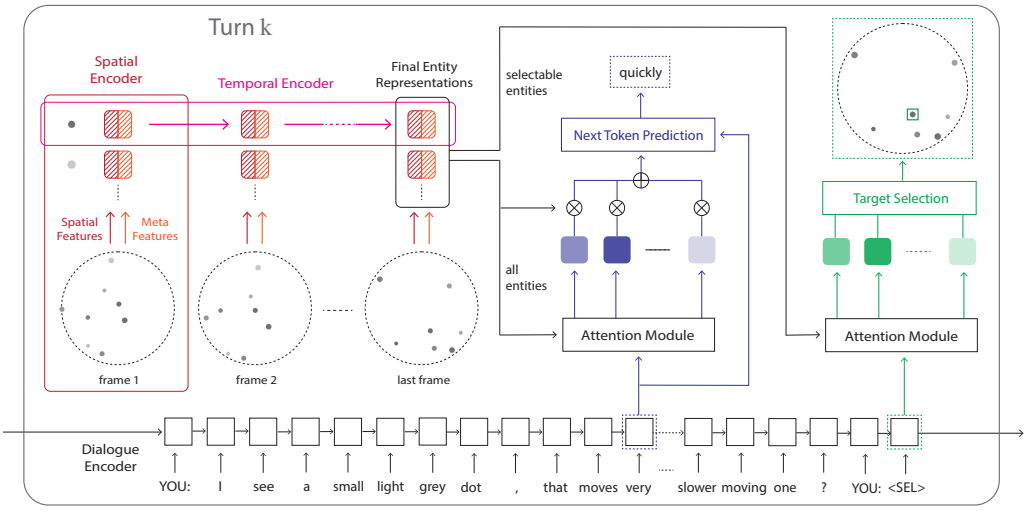



Figure 6.6: Our baseline model architecture. Information flow in turn $k$ is illustrated. When generating model utterances (in selfplay dialogue and human evaluation), we sample next tokens with the temperature set to 0.25.

computed as:

$$\mathrm{MLP}(\mathbf{e}_i) \odot \sum_{\substack{e_j \in \tilde{E}, \\ j \neq i}} \mathrm{MLP}(\mathbf{e}_i - \mathbf{e}_j) \tag{6.1}$$

where $\mathbf{e}_i$ is the vector representation of entity $e_i$ and $\odot$ is the vector concatenation.[10]

Meta features are binary information of each entity representing whether (or not) the entity (i) is visible in the frame, (ii) is visible at timestep $t_k$, (iii) was visible at timestep $t_{k-1}$, and (iv) was selected in the previous turn (i.e. is the *previous target*). Meta features are also encoded using an MLP, and we take the sum of spatial/meta features as the (entity-level) output of the spatial encoder.

Finally, we use the temporal encoder based on a GRU to encode the outputs of the spatial encoder. The final state of the temporal encoder is considered as the final representation of each entity.

Based on the outputs of these encoders, we use two attention modules (based on MLPs) to compute attention scores for each entity. The first attention module is used to weight the final representations of all entities $E_{all}$ conditioned on the current dialogue state: then, the weighted sum of $E_{all}$ is concatenated with the dialogue state to predict the next dialogue token (Xu et al., 2015). The second module is used to predict the target entity, where we simply take the (soft)max of attention scores for the selectable entities $E_{sel}$ in turn $k$.

Note that there are only two main differences between our baseline and the best OCC model (TSEL-REF-DIAL) from Udagawa and Aizawa (2020): first, in TSEL-REF-DIAL, the final representation of each entity is its *spatial features*, i.e. the meta features and temporal encoder are not used (which are only meaningful in D-OCC). Second, TSEL-REF-DIAL is also trained on the *reference resolution* task (using an additional attention module), which is only available in OCC. Due to this architectural similarity, we can virtually *pretrain* our model on OCC by initializing the shared model parameters based on TSEL-REF-DIAL and then fine-tune the whole model on D-OCC.[11]

[10] To be precise, $\mathbf{e}_i$ is a 4-dimensional vector representing color, size, and 2-D location. If the entity is not observable in the frame, we use the default value of (0, 0) for the location.

[11] For pretraining, we retrained TSEL-REF-DIAL with the shared word embedding for OCC and D-OCC.

### 6.5.3 Experiment Setup

All modules of our baseline (MLPs and GRUs) are single layered with 256 hidden units, except for the attention modules which are 2-layered. Dropout rate of 0.5 is applied at each layer during training, and we use the Adam optimizer (Kingma and Ba, 2015) with the initial learning rate set to 0.001. After manual tuning on the validation set, we weight the losses from next token prediction and target selection with the ratio of 2:1.

In terms of data splits, we use 500 dialogues with LST $\geq 2$ for testing target selection, another 500 for validation and the rest for training.[12] Note that we use all unsuccessful turns (where the players failed to agree upon the same entity) as well, assuming they are still based on valid strategies. For selfplay dialogue and human evaluation, we collect 2,000 and 200 dialogues in unseen environments, respectively. Each experiment is repeated 5 times with different random seeds (including data splits), except for human evaluation.

Finally, we conduct extensive ablations to study the effect of various model architectures, including *pretraining*, spatial attributes (*color*, *size* and *location*) and the meta feature (*previous target*). In addition, we also ablate the *dynamic* information of the observation by only using the last frame in each turn as the input for the temporal encoder.

### 6.5.4 Results

| Model | Turn / Previous Target | | |
|---|---|---|---|
| | 1st / - | ≥2nd / ✓ | ≥2nd / ✗ |
| Baseline | **76.4**±**1.7** | 96.6±0.3 | 67.4±0.5 |
| − pretraining | 74.6±2.7 | 96.3±0.7 | 66.9±1.1 |
| − color | 56.3±2.0 | 95.7±0.6 | 50.5±1.4 |
| − size | 58.4±1.3 | 95.7±0.9 | 52.2±0.5 |
| − location | 74.4±1.5 | 96.1±0.9 | 67.3±0.7 |
| − previous target | 76.1±1.7 | 83.3±1.1* | **67.8**±**0.6*** |
| − dynamics | 75.1±2.3 | **96.7**±**1.0** | 67.0±0.7 |
| Human | 97.0±1.1 | 98.2±0.5* | 95.8±2.0* |

Table 6.6: Results for the target selection task (* denotes cases where the correct previous targets were not provided during prediction).

We show the results for target selection in Table 6.6. The human performance is estimated by 3 annotators based on 50 dialogues with LST $\geq 2$.

Based on these results, we can verify that all ablations hurt the performance of our baseline in some way. Pretraining on OCC is generally effective, and all spatial attributes contribute to the overall performance (especially color and size). When the meta feature of the correct previous target is available, all models perform remarkably well in ✓ cases (previous target stays in common), which is natural since humans often repeated the same selection. Finally, dynamic information also contributes to the baseline performance, despite the effect being rather marginal.

However, there is huge room left for improvement in the 1st turn and even more so in ✗ cases (previous target no longer in common). These results indicate that recognizing the *creation* of common ground is still difficult, and recognizing how they are *updated* (rather than *retained*) remains even more challenging for the current baseline.

Next, we show the results for selfplay dialogue and human evaluation in Tables 6.7 and 6.8, respectively. We also include the results of TSEL-REF-DIAL (trained on OCC

[12] We ensured no overlaps in terms of the *environments* across data splits.

| | | | | **Seflplay Dialogue** | | | |
|---|---|---|---|---|---|---|---|
| Model | Dataset | Turn | Previous Target | Success Rate (%) #Shared=4 | #Shared=5 | #Shared=6 | Avg. LST |
| Baseline | D-OCC | 1st | - | 46.8±1.8 | 63.8±1.8 | 80.2±2.3 | |
| | | ≥2nd | ✓ | 99.4±0.3 | 99.7±0.2 | 99.6±0.2 | 1.94±0.09 |
| | | | ✗ | **48.5±2.2** | **64.6±2.8** | **81.5±1.5** | |
| − pretraining | D-OCC | 1st | - | 39.4±1.0 | 53.5±0.8 | 73.7±1.8 | |
| | | ≥2nd | ✓ | 98.6±2.4 | 98.8±1.8 | 99.4±1.0 | 1.35±0.09 |
| | | | ✗ | 30.3±5.7 | 42.1±6.3 | 65.4±4.9 | |
| − color | D-OCC | 1st | - | 36.3±2.0 | 54.6±2.3 | 72.9±1.5 | |
| | | ≥2nd | ✓ | 99.7±0.1 | 99.7±0.0 | 99.6±0.1 | 1.50±0.10 |
| | | | ✗ | 42.1±3.5 | 56.7±4.2 | 72.4±4.6 | |
| − size | D-OCC | 1st | - | 41.5±0.8 | 58.0±0.9 | 75.2±1.3 | |
| | | ≥2nd | ✓ | 99.8±0.1 | 99.7±0.1 | 99.8±0.2 | 1.58±0.07 |
| | | | ✗ | 39.6±3.5 | 55.3±3.6 | 69.9±1.5 | |
| − location | D-OCC | 1st | - | 45.7±1.9 | 60.4±1.6 | 77.7±1.7 | |
| | | ≥2nd | ✓ | 99.8±0.1 | 99.7±0.0 | 99.7±0.1 | 1.68±0.09 |
| | | | ✗ | 40.8±3.6 | 54.6±2.5 | 73.9±4.2 | |
| − previous target | D-OCC | 1st | - | **49.2±1.3** | 64.0±1.8 | 82.2±2.0 | |
| | | ≥2nd | ✓ | 85.8±2.7 | 87.5±1.6 | 91.2±1.3 | 1.45±0.05 |
| | | | ✗ | 29.2±1.5 | 41.9±1.9 | 64.5±1.0 | |
| − dynamics | D-OCC | 1st | - | **49.2±2.2** | **65.8±1.3** | **83.3±1.9** | |
| | | ≥2nd | ✓ | **99.9±0.1** | **99.9±0.1** | **99.8±0.1** | **2.02±0.07** |
| | | | ✗ | 48.3±2.2 | 63.5±2.8 | 81.1±2.1 | |
| TSEL-REF-DIAL | D-OCC | 1st | - | 41.0±1.2 | 58.7±1.1 | 76.0±1.8 | - |
| | OCC | 1st | - | 45.9±1.6 | 62.7±2.2 | 79.7±1.0 | |
| Human | D-OCC | 1st | - | 73.4 | 82.0 | 87.6 | |
| | | ≥2nd | ✓ | 95.4 | 97.0 | 97.8 | 3.31 |
| | | | ✗ | 81.7 | 88.4 | 91.6 | |

Table 6.7: Results for the selfplay dialogue task. Human performance is estimated based on the overall average of the crowd workers (c.f. Table 6.1 and 6.4).

| | | | | **Human Evaluation** | |
|---|---|---|---|---|---|
| Model | Dataset | Turn | Previous Target | Success Rate (%) | Avg. LST |
| Baseline | D-OCC | 1st | - | **44.5** | |
| | | ≥2nd | ✓ | 81.9 | **1.00** |
| | | | ✗ | **44.4** | |
| − location | D-OCC | 1st | - | 40.0 | |
| | | ≥2nd | ✓ | **91.8** | 0.81 |
| | | | ✗ | 36.3 | |
| − dynamics | D-OCC | 1st | - | 37.0 | |
| | | ≥2nd | ✓ | 86.8 | 0.79 |
| | | | ✗ | 39.2 | |
| Human | D-OCC | 1st | - | 80.5 | |
| | | ≥2nd | ✓ | 96.7 | 3.31 |
| | | | ✗ | 86.6 | |

Table 6.8: Results for the human evaluation. Human performance is estimated based on the overall average of the crowd workers (c.f. Table 6.1 and 6.4).

*without* fine-tuning on D-OCC) as a reference.[13]

In selfplay dialogue, we can verify that the baseline model performs reasonably well, outperforming TSEL-REF-DIAL in the 1st turn of D-OCC (as well as OCC). However, it is worth noting that TSEL-REF-DIAL may be suffering from a minor covariate shift in D-OCC (c.f. Section 6.2.2), and without pretraining, our baseline still underperforms this best OCC model. We also found that all ablations of spatial attributes hurt perfor-

[13]When testing TSEL-REF-DIAL on D-OCC, we used the spatial features of the last observation frame as the input.

| Dataset | Context Type | | | Context Update | Context Source | Evaluation of Common Grounding |
|---|---|---|---|---|---|---|
| | CNT | P.O. | DYN | | | |
| Twitch-FIFA (Pasunuru and Bansal, 2018) | ✓ | ✗ | ✓ | ✓ | Synthetic | N/A |
| AVSD (Alamri et al., 2019) | ✓ | ✓ | ✓ | ✗ | Real | Indirect |
| SIMMC (Moon et al., 2020) | ✓ | ✗ | ✗ | ✓ | Synthetic+Real | Indirect |
| MutualFriends (He et al., 2017) | ✗ | ✓ | ✗ | ✗ | Synthetic | Create |
| GuessWhat?! (De Vries et al., 2017) | ✓ | ✗ | ✗ | ✗ | Real | Create |
| Photobook Dataset (Haber et al., 2019) | ✓ | ✓ | ✗ | ✓ | Real | Create |
| OneCommon (Udagawa and Aizawa, 2019) | ✓ | ✓ | ✗ | ✗ | Synthetic | Create |
| **Dynamic-OneCommon (Ours)** | ✓ | ✓ | ✓ | ✓ | **Synthetic** | **Create+Maintain** |

Table 6.9: Comparison with the major datasets. Context type is considered *dynamic* if it involves rich, spontaneous dynamics (as in videos) and contexts to be *updated* if new information is provided in the course of the dialogue (CNT = *continuous*, P.O. = *partially-observable*, DYN = *dynamic*).

mance, while the locational attributes became more critical in the full dialogue task. The meta feature of the previous target (selected by the model) is also critical, as the models seem to be relying heavily on this feature to both retain and update the target.

However, we found that ablation of dynamic information does not degrade (actually improves) performance in selfplay dialogue. This indicates that the last frame of each turn (*current state*) is sufficient for the baseline to coordinate with itself, and it is unlikely to be leveraging sophisticated temporal information (*state change* or *previous state*) like the human strategies seen in Section 6.4.2. Also, while the models perform near perfectly in ✓ cases, the success rates drop or do not improve significantly in ✗ cases (compared to the 1st turn). This shows that current models can *retain* the same common ground easily but struggle in *updating* them using the previous common ground, unlike the human strategies seen in Section 6.4.3.

Finally, in human evaluation, we could verify that our baseline performs the best of the top 3 models in the selfplay dialogue task, but the success rates were much lower than observed in selfplay. This indicates that current models may not be using natural language in the same way humans use it (i.e. are not properly *grounded*, Bender and Koller, 2020), although they do become closer to it when all the features are available.[14]

To summarize, our results in sequential collaborative reference show that the current baseline can leverage all spatial features and retain the same common ground, especially when provided explicitly as the meta feature. However, it may not be using temporal information effectively, and the creation and update of common ground still remain challenging in the dynamic environments, especially when conversing with real humans.

## 6.6 Related Work

The notion of common ground was originally introduced in Lewis (1969) and Stalnaker (1978) and theoretically elaborated in fields such as psycholinguistics (Clark and Brennan, 1991; Brennan et al., 2010). While formal approaches (rule/logic-based) exist to computationally model the process of common grounding (Traum, 1994; Van Ditmarsch et al., 2007; Poesio and Rieser, 2010), capturing their full complexities in realistic, situated conversations remains a formidable problem.

From an empirical perspective, various dialogue tasks have been proposed to develop and evaluate data-driven models of common grounding. Most of the existing literature focuses on closed domain, goal-oriented settings to measure the ability both quantitatively and objectively (Fang et al., 2015; Zarrieß et al., 2016; De Vries et al., 2017).

[14] At the *superficial* level, all models could generate fluent utterances and complete the task with minimal confusion.

Recent works, summarized as the *grounded agreement games* in Schlangen (2019), introduce symmetric speaker roles to encourage more bilateral interaction. Udagawa and Aizawa (2019) also raise *continuous* and *partially-observable* context to be essential for requiring advanced common grounding (Section 6.2.1). Finally, Haber et al. (2019) propose a multi-round image identification task, where different combinations of images are provided to each agent at every round. While this setting is useful for studying *subsequent references* affected by the existing common ground (Brennan and Clark, 1996; Takmaz et al., 2020), the observations in each round are static, temporarily independent images. Hence, all of these tasks focus on creating common ground under *static* context and lack evaluation metrics for *maintaining* common ground in dynamic environments.

We also note that some recent dialogue tasks require dealing with dynamic information, although common grounding usually takes place *implicitly* and may be difficult to measure directly. For instance, Alamri et al. (2019) proposed Q&A based dialogues grounded in video contexts. However, the information given to each agent remains fixed throughout the dialogue, requiring *creation* but minimal *update* of common ground. Many recent works also focus on dialogues grounded in external environments (de Vries et al., 2018; Suhr et al., 2019; Narayan-Chen et al., 2019; Thomason et al., 2019; Moon et al., 2020). These settings often involve dynamic change of the *perspectives*, but they usually assume the environments themselves to be stationary and do not change spontaneously (without direct intervention). In contrast to these works, we introduce both *context updates* in the course of the dialogue and *diverse dynamics* of the external environment to require advanced common grounding.[15] We summarize our comparison with the major existing datasets in Table 6.9.

Finally, our work is relevant to the emerging literature on spatio-temporal grounding in computer vision and NLP. This includes video QA (Lei et al., 2018; Yu et al., 2019b; Castro et al., 2020), video object grounding (Zhou et al., 2018; Chen et al., 2019b; Sadhu et al., 2020) and video captioning (Krishna et al., 2017a), all of which are essential subtasks in our dialogue. However, existing resources often contain exploitable biases and lack visual/linguistic diversity as well as reliable evaluation metrics (esp. in language generation) (Aafaq et al., 2019). It is also challenging to probe model behaviors without the controllability of the video contexts (Girdhar and Ramanan, 2020). We have addressed such concerns based on our task design (Section 6.2.2) and expect our resource to be useful for promoting this line of research as well.

## 6.7 Discussion and Conclusion

In this study, we proposed a novel dialogue task to study the ability of creating, retaining and updating common ground in dynamic environment. The dynamics of the environment are fully controllable in our setting, which allows us to introduce diverse strategies while making exploitation of dataset biases difficult. Based on our dataset analyses and experiments, we demonstrated the advanced strategies of common grounding required and the open room for improvement in our newly developed Dynamic-OneCommon Corpus (D-OCC).

D-OCC can be utilized and enriched in several ways. For instance, we can conduct various *causal analysis*, e.g. by changing certain feature of entities (such as movement) and studying the differences in model behavior, which is essential yet difficult to conduct in many existing datasets (c.f. Section 6.6). We can also add fine-grained annotation of *reference resolution* (Chapter 4), as (partially) illustrated in Figure 6.1. Finally, we

[15] While Pasunuru and Bansal (2018) collected live-stream dialogues grounded in soccer video games, the non-goal-oriented, unconstrained nature of their setting makes evaluation and analysis of common grounding very challenging.

can annotate *spatio-temporal expressions*, e.g. by following the procedure discussed in Chapter 5. Such annotations would allow us to gain deeper understandings of the *intermediate* process of common grounding: for instance, we can study whether the developed models recognize and use the spatio-temporal expressions appropriately and consistently in a *human-like* way (i.e. not only imitate at the superficial level, as observed in Section 6.5.4).

We'd also like to discuss the main limitation of our current work, namely the *ecological validity* (De Vries et al., 2020) of D-OCC. Since we focused on the simplest task setting under *continuous*, *partially-observable* and *dynamic* context, direct application of our work in realistic settings may not be straightforward. However, the rich variety of spatio-temporal expressions observed in our synthetic environment is fundamental in many real-world settings (e.g. locating entities in the crowd, traffic or disaster scenes, Pustejovsky et al., 2011a), and we expect D-OCC to be essential for developing and diagnosing various models for generic spatio-temporal grounding. In addition, our sequential collaborative reference task is defined generally (Section 6.2.2), and we can easily scale up the task complexity to study common grounding under the desired dynamics in consideration.

Overall, we expect our task design, resource and analyses to be fundamental for developing dialogue systems that can both create and maintain common ground reliably in dynamic environments.

# Chapter 7

# Discussion on Future Research

As a final discussion, we explore the future prospects of our research. Specifically, we present further ideas on the *task design methodologies* to study fully advanced common grounding (Section 7.1), how to further *improve common grounding* in existing dialogue systems (Section 7.2), and the implication of our contributions on *real-world applications* (Section 7.3).

## 7.1 Task Design Methodologies

So far, we have focused on the (sequential) collaborative reference tasks, which require coordination at the level of mutual entity recognition (i.e. *entity-level alignment*). To represent the complexity of situated common grounding, we incorporated three universal factors of realistic environments into the dialogue context: namely *continuity*, *partial-observability* and *dynamics*. We've empirically investigated how each factor introduces complexity of common grounding and why they need to be taken into account in the dialogue task design.

While our contributions are *necessary* to study advanced common grounding, they are still not *sufficient* to embrace all aspects of full-fledged human common grounding. For instance, mutual entity recognition is only the first step of general common grounding, and we will need task formulations which enable accurate evaluation of common grounding in its entirety. Moreover, to fully replicate the complexity of common grounding, additional factors of real-world settings need to be taken into account: such as those related to physical commonsense and psychological reasoning.

To this end, we expect that the (relatively overlooked) view of Pickering and Garrod (2004) will play a key role. As discussed in Section 2.1.1, this view considers common grounding as the alignment of *situation models* among the interlocutors. To be precise, the mental representations of the state of affairs become aligned through the process of common grounding in several dimensions: notably *space*, *time*, *causality*, *intentionality*, and *protagonists* (Zwaan and Radvansky, 1998).

Consider the case of common grounding with regard to a football game. At the very least, representations in the *space* and *time* dimensions need to be aligned, e.g. the temporal sequence of salient events (related to the movements of the ball and players). In addition, representations must be aligned in the *causality* dimension, i.e. the evident causal chains between such events. Furthermore, the *intentionality* dimension involves the understanding of each player's (or even the team's) goal and plans, potentially requiring the domain knowledge of football strategies. Finally, the *protagonists* dimension may be concerned with the precise representation of each player, including his/her position, style of play, condition, background profile, and so forth.

Based on this view, our proposed tasks are mostly focused on the primary dimensions of *space* and *time*. In OneCommon Corpus (Chapter 3), each agent starts with a *spatial*

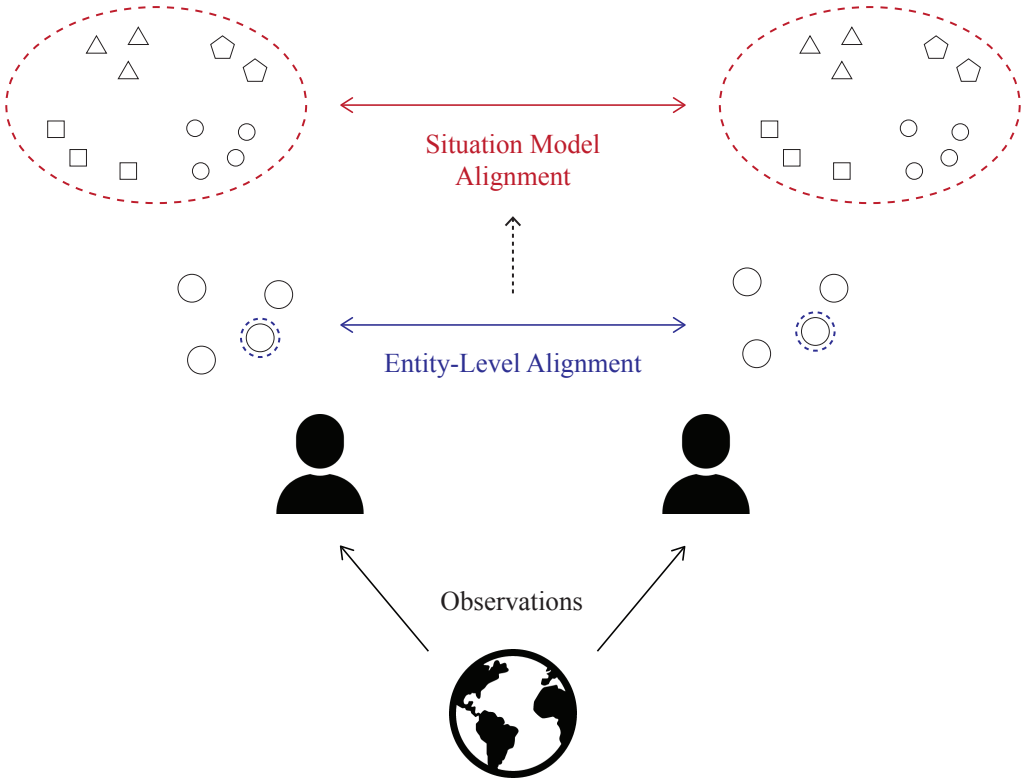



Figure 7.1: An illustration of the current task formulations (*entity-level alignment*) and the desired task formulations (*situation model alignment*).

situation model and tries to align this with the partner by identifying the same, common entity. In Dynamic-OneCommon Corpus (Chapter 6), each agent holds a stream of *spatio-temporal* situation models, and their goal is to keep these models aligned by identifying the same entities at certain intervals. While the accuracy of common ground is easy to measure in these settings (based on the success rates in collaborative reference), they do not reflect the alignment of the interlocutors' entire situation models.

If we were to study fully advanced common grounding, we expect task formulations which require the alignment of entire situation models will be critical (as illustrated in Figure 7.1). To be specific, we need appropriate task designs which require alignment at all conceivable dimentions (not only *space* and *time* but also *causality*, *intentionality* and *protagonists*). We also need task formulations which enable reliable evaluation of the accurate situation model alignment (rather than entity-level alignment). These requirements are much more demanding, but there are several potential solutions. One approach is to carefully design questions to test whether the situation models are correctly aligned, as similarly proposed in the reading comprehension literature (Sugawara et al., 2021). For instance, after (or in the middle of) the conversation, we can ask each interlocutor some questions that require the answers to be coordinated: e.g. "what event do you (and your partner) expect if the player had not scored an own goal?" or "what do you (and your partner) believe the intention of the player was?", where alignment of situation models at the *causal* (*counterfactual*) dimension and *intentionality* dimension is required, respectively. We leave the exploration of specific task designs, evaluation metrics and dataset collection as an important avenue of future work.

## 7.2 Improving Common Grounding

Throughout this thesis, we have proposed several approaches to improve common grounding in data-driven dialogue systems: such as multi-task training with reference resolution (Chapter 4) and pretraining on a related dataset (Chapter 6). However, we can conceive of three major approaches to further improve their performances.

One approach is to make the model learn from the task success and failure based on *reinforcement learning* (Sutton and Barto, 2018). This can be realized through an actual interaction with the human players (e.g. crowd workers) but also by repeatedly playing the task against itself, known as *selfplay* (Silver et al., 2016). The latter approach is much cheaper and more scalable, hence widely applied in dialogue domains with symmetric speaker roles (Lewis et al., 2017; Yarats and Lewis, 2018; Jang et al., 2020). The upside is that we can expect the models to learn to avoid ineffective strategies, such as underspecification and premature guessing. However, the downside is that they can potentially diverge from natural language (Kottur et al., 2017), making communication with humans more difficult and unreliable.

We also expect the incorporation of *pragmatic reasoning* to be a fruitful area of future research. One representative approach is the *Rational Speech Act* (*RSA*) framework (Goodman and Frank, 2016), which has been applied in both continuous (Monroe et al., 2017; McDowell and Goodman, 2019) and partially-observable domains (Hawkins et al., 2021). However, application in *interactive* and *dynamic* domains would involve additional complexities that need to be taken into account, such as the dependencies on dialogue history and previous common ground.

Finally, we can study wider variety of model architectures and pretraining datasets. For instance, we have considered the observations to be in *structured* format (e.g. based on xy coordinates), but we can also extract *raw* visual features by directly treating them as images or videos (Suhr et al., 2017; Iki and Aizawa, 2020). In the latter case, we can apply various techniques from CV and NLP, including image/video processing methods (Carreira and Zisserman, 2017; Wang et al., 2018b; Dosovitskiy et al., 2021), vision-language grounding models (Lu et al., 2019b; Le et al., 2020) and pretraining on large-scale, open domain datasets (Krishna et al., 2017b; Sharma et al., 2018). Note that the entity-level representation of the observation (required in our baselines) can be obtained from raw visual features as well, e.g. by utilizing the object detectors (Ren et al., 2015; Redmon et al., 2016) and trackers (Bergmann et al., 2019; Wang et al., 2020). Lastly, we can replace the *dialogue encoders* of our baselines based on the transformer architectures (Vaswani et al., 2017), which have become the mainstream approch in language and dialogue modelling (Devlin et al., 2019; Lewis et al., 2020; Zhang et al., 2020).

## 7.3 Real-World Applications

Finally, we'd like to discuss the impact of our work on real-world applications. While our resources (OneCommon Corpus and Dynamic-OneCommon Corpus) are currently restricted to synthetic settings, the general ideas and strategies we have investigated are fundamental in many practical applications.

For instance, consider the case of *item retrieval*, such as recommending a furniture by soliciting the user preferences (as in Moon et al., 2020). Since the user's taste can be very precise, this requires dealing with subtle nuances like "do you have something that's *a little more* white and round shaped?" or partial replacements like "I like its materials *except for* the backrest". Dealing with such expressions are critical under *continuous* context and can be studied rigorously based on OneCommon Corpus. In addition, this setting can become *partially-observable* in case the actual furniture is not visible to the user, e.g. when the item is out of stock or the user is talking to speech-

based devices. On such occasions, the system needs to take into account the possibility of various misunderstandings to make the transaction accurate and reliable.

We can also illustrate the importance of *dynamic* context based on a navigation task in commonplace environments, such as finding a lost child in an urban city. In reality, the *target entity* (the child) may not stay in one place, so the routing directions can no longer be fixed and need to be *updated* accordingly (as in "now head more to the west" or "go back to the previous block"). Furthermore, the *landmark entities* may not be stationary either and could be ephemeral (as in "following the group of travelers" or "in the middle of the crowd"). The task may be trivial if the child is conspicuous with few distractors, but otherwise (e.g. with many pedestrians around), the descriptions need to be precise and distinguishing (as in "wearing *a little* darker shirt", "walking *right* towards the station"). In order to study such (nuanced and pragmatic) spatio-temporal expressions and references to previous common ground, we expect Dynamic-OneCommon Corpus to be an essential proving ground for developing and analyzing various models.[1]

Overall, we expect our contributions to be critical for promoting reliable collaboration in real-world environments, which involve the advanced settings of *continuous*, *partially-observable* and *dynamic* context.

[1]Unfortunately, most of the existing navigation tasks focus on *static* environments (e.g. static targets and landmarks): see Section 6.6 for further discussions.

# Chapter 8

# Conclusion

Developing systems that can understand and communicate in natural language is an ultimate goal of NLP and AI research. In this thesis, we focused on the aspect of common grounding as the key requisite for truly reliable conversation.

In Chapter 2, we gave an overview of the existing research related to common grounding. First, we introduced the theoretical foundations of common grounding in philosophy, mathematical logic and pyscholinguistics. Then, we explained the computational approaches to common grounding, including the general literature on dialogue system engineering. Finally, we discussed the important links to the problem of symbol grounding, which is deeply related to common grounding (especially in situated dialogues). Overall, we clarified the limitations of existing research and emphasized the importance of our research in these broader contexts.

In Chapter 3, we introduced a novel task setting under continuous and partially-observable context. Following this idea, we designed a minimal collaborative reference task and developed OneCommon Corpus containing 6,760 dialogues. Through our dataset analysis, we verified advanced common grounding required in this setting, such as the collaborative resolution of complex ambiguity, uncertainty and misunderstandings. As a preliminary experiment, we evaluated a simple baseline model on the target selection task and verified the difficulty of even recognizing the common ground.

In Chapter 4, we proposed a method of decomposing common grounding based on its subtask of reference resolution. Based on our simple and generic framework, we annotated 5,191 successful dialogues from OneCommon Corpus which capture genuine ambiguity while maintaining reliability. Our dataset analysis demonstrated the importance of our annotation for interpreting the intermediate process of common grounding. Finally, we built end-to-end dialogue systems as the baselines for our proposed dataset. Our annotation helped improve and interpret their common grounding strategies, but substantial room remained for further improvement.

In Chapter 5, we conducted further analyses of OneCommon Corpus based on spatial expressions. Specifically, we leveraged the existing annotation of referring expressions to annotate their spatial predicates (including modifications) efficiently and reliably. Although our annotation size is relatively small (600 dialogues), we revealed important linguistic structures in our dataset, i.e. predicate-argument structure, modification and ellipsis. Finally, through our extensive experiments, we showed how and where the current baselines struggle in capturing such precise structures.

In Chapter 6, we further take into account the task setting under dynamic environments. To evaluate and analyze the ability of mainining common ground, we formulated the sequential collaborative reference task and collected Dynamic-OneCommon Corpus containing 5,617 dialogues. Our dataset analyses demonstrated sophisticated strategies required in this setting, such as the usage of spatio-temporal expressions and references to previous common ground. In our experiments, we conducted thorough evaluation of

the end-to-end dialogue systems based on this dataset. Overall, we showed that there remains significant room left for improvement due to the requirement of even more advanced common grounding strategies.

In Chapter 7, we briefly explored the future prospects of our research. In terms of the task design methodologies, we expect the view of Pickering and Garrod (2004) will be crucial for the next-generation task formulations reflecting fully advanced common grounding. Next, we gave an overview of the promising directions to further improve the model performances in terms of common grounding. Finally, we illustrated several potential applications of our contributions in the practical, real-world settings.

All together, we expect this thesis to be a fundamental basis for realizing truly reliable communication between humans and computers. All of our resources are publicly available to facilitate future research, and we hope they will encourage further analyses and improvements of dialogue systems in terms of advanced common grounding.

# References


Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Comput. Surv.*, 52(6), October 2019. ISSN 0360-0300.

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*, 2020.

Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. History for visual dialog: Do we really need it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8182–8197, Online, July 2020. Association for Computational Linguistics.

C. Aggarwal and ChengXiang Zhai. Mining text data. In *Springer US*, 2012.

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7558–7567, 2019.

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. The hcrc map task corpus. *Language and speech*, 34(4):351–366, 1991.

Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, Hao Fang, Alan Guo, David Hall, Kristin Hayes, Kellie Hill, Diana Ho, Wendy Iwaszuk, Smriti Jha, Dan Klein, Jayant Krishnamurthy, Theo Lanman, Percy Liang, Christopher H. Lin, Ilya Lintsbakh, Andy McGovern, Aleksandr Nisnevich, Adam Pauls, Dmitrij Petters, Brent Read, Dan Roth, Subhro Roy, Jesse Rusak, Beth Short, Div Slomin, Ben Snyder, Stephon Striplin, Yu Su, Zachary Tellman, Sam Thomson, Andrei Vorobev, Izabela Witoszko, Jason Wolfe, Abby Wray, Yuchen Zhang, and Alexander Zotov. Task-oriented dialogue as dataflow synthesis. *Transactions of the Association for Computational Linguistics*, 8:556–571, 2020.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

Nicholas Asher, Nicholas Michael Asher, and Alex Lascarides. *Logics of conversation*. Cambridge University Press, 2003.

John Langshaw Austin. *How to do things with words*. Oxford university press, 1962.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. Phyre: A new benchmark for physical reasoning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 5082–5093. Curran Associates, Inc., 2019.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pages 2670–2676, 2007.

Yonatan Belinkov and James Glass. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 2019.

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

Luciana Benotti and Patrick Blackburn. Grounding as a collaborative process. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 515–531, Online, April 2021. Association for Computational Linguistics.

Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. Tracking without bells and whistles. In *International Conference on Computer Vision*, 2019.

Pierre Bézier. Mathematical and practical possibilities of UNISURF. In *Computer Aided Geometric Design*, pages 127–152. Elsevier, 1974.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online, November 2020. Association for Computational Linguistics.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. Learning end-to-end goal-oriented dialog. In *International Conference on Learning Representations*, 2017.

S. Brennan and H. H. Clark. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology. Learning, memory, and cognition*, 22 6:1482–93, 1996.

Susan E. Brennan. The grounding problem in conversations with and through computers. In S. R. Fussell and R. J. Kreuz, editors, *Social and Cognitive Approaches to Interpersonal Communication*, pages 201–225. Lawrence Erlbaum, Hillsdale, NJ, 1998.

Susan E Brennan, Alexia Galati, and Anna K Kuhlen. Two minds, one dialog: Coordinating speaking and understanding. In *Psychology of learning and motivation*, volume 53, pages 301–344. Elsevier, 2010.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

Harry Bunt, Volha Petukhova, David Traum, and Jan Alexandersson. Dialogue act annotation with the iso 24617-2 standard. In *Multimodal Interaction with W3C Standards*, pages 109–135. Springer, 2017.

J. Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4733, 2017.

Santiago Castro, Mahmoud Azab, Jonathan Stroud, Cristina Noujaim, Ruoyao Wang, Jia Deng, and Rada Mihalcea. LifeQA: A real-life dataset for video question answering. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4352–4358, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4.

Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, and Devi Parikh. Evaluating visual conversational agents via cooperative human-ai games. In *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2017.

Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh. Attacking visual language grounding with adversarial examples: A case study on neural image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2587–2597, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. Towards knowledge-based recommender dialog system. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1803–1813, Hong Kong, China, November 2019a. Association for Computational Linguistics.

Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee Kenneth Wong. Weakly-supervised spatio-temporally grounding natural sentence in video. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1884–1894, Florence, Italy, July 2019b. Association for Computational Linguistics.

Hyundong Cho and Jonathan May. Grounding conversations with improvised dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2398–2413, Online, July 2020. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111. Association for Computational Linguistics, 2014.

Noam Chomsky. *Syntactic Structures*. Mouton, 1957.

Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. Visual referring expression recognition: What do systems actually learn? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 781–787, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Eve V Clark and James B Grossman. Grounding and attention in language acquisition. In *Papers from the 37th meeting of the Chicago Linguistic Society*, volume 1, pages 95–116, 2001.

H. Clark and Susan E. Brennan. Grounding in communication. In *Perspectives on socially shared cognition*, 1991.

Herbert H Clark. *Using language*. Cambridge university press, 1996.

Herbert H. Clark and Catherine R. Marshall. Definite knowledge and mutual knowledge. In Aravind K. Joshi, Bonnie L. Webber, and Ivan A. Sag, editors, *Elements of Discourse Understanding*, pages 10–63. Cambridge, UK: Cambridge University Press, 1981.

Herbert H Clark and Edward F Schaefer. Contributing to discourse. *Cognitive science*, 13(2):259–294, 1989.

Jacob Cohen. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

David A Cook and Thomas J Beckman. Current concepts in validity and reliability for psychometric instruments: theory and application. *The American journal of medicine*, 119(2):166–e7, 2006.

Soham Dan, Parisa Kordjamshidi, Julia Bonn, Archna Bhatia, Zheng Cai, Martha Palmer, and Dan Roth. From spatial relations to spatial configurations. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5855–5864, Marseille, France, 2020. European Language Resources Association. ISBN 979-10-95546-34-4.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 2017a.

Abhishek Das, Satwik Kottur, Jose M. F. Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017b.

Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017c.

Sam Davidson, Dian Yu, and Zhou Yu. Dependency parsing for spoken dialog systems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1513–1519, Hong Kong, China, November 2019. Association for Computational Linguistics.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proc. of CVPR*, 2017.

Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. Talk the walk: Navigating new york city through grounded dialogue. *arXiv preprint arXiv:1807.03367*, 2018.

Harm De Vries, Dzmitry Bahdanau, and Christopher Manning. Towards ecologically valid research on language user interfaces. *arXiv preprint arXiv:2007.14435*, 2020.

Jan Deriu, Álvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, S. Rosset, Eneko Agirre, and Mark Cieliebak. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755 – 810, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. Towards end-to-end reinforcement learning of dialogue agents for information access. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–495, Vancouver, Canada, July 2017. Association for Computational Linguistics.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin D. Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. Frames: A corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219. Association for Computational Linguistics, August 2017.

Desmond Elliott and Frank Keller. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

Jeffrey L Elman, Elizabeth A Bates, and Mark H Johnson. *Rethinking innateness: A connectionist perspective on development*, volume 10. MIT press, 1996.

Ronald Fagin, Yoram Moses, Joseph Y Halpern, and Moshe Y Vardi. *Reasoning about knowledge*. MIT press, 2003.

Rui Fang, Malcolm Doering, and Joyce Y. Chai. Embodied collaborative referring expression generation in situated human-robot interaction. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 271–278, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-2883-8.

Erika L Ferguson and Mary Hegarty. Properties of cognitive maps constructed from texts. *Memory & cognition*, 22(4):455–473, 1994.

John R Firth. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1957.

Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

J. Fodor and Z. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28:3–71, 1988.

Zhe Gan, Yu Cheng, Ahmed Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. Multi-step reasoning via recurrent dual attention for visual dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6463–6474, Florence, Italy, July 2019. Association for Computational Linguistics.

Jianfeng Gao, Michel Galley, and Lihong Li. Neural approaches to conversational ai. *Foundations and Trends® in Information Retrieval*, 13(2-3):127–298, 2019. ISSN 1554-0669.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating nlp models via contrast sets. *arXiv preprint arXiv:2004.02709*, 2020.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Mor Geva, Yoav Goldberg, and Jonathan Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China, November 2019. Association for Computational Linguistics.

Jonathan Ginzburg. *The interactive stance*. Oxford University Press, 2012.

Rohit Girdhar and Deva Ramanan. CATER: A diagnostic dataset for Compositional Actions and TEmporal Reasoning. In *ICLR*, 2020.

Arthur M Glenberg, Marion Meyer, and Karen Lindem. Mental models contribute to foregrounding during text comprehension. *Journal of Memory and Language*, 26(1): 69–83, 1987. ISSN 0749-596X.

Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016.

Noah D. Goodman and Michael C. Frank. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20:818–829, 2016.

Jana Götze and Johan Boye. SpaceRef: A corpus of street-level geographic descriptions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3822–3827, Portorož, Slovenia, 2016. European Language Resources Association (ELRA).

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.

H Paul Grice. Meaning. *The philosophical review*, 66(3):377–388, 1957.

H. Paul Grice. Logic and conversation. *Syntax and Semantics*, 3:41–58, 1975.

N. Gupta, Gökhan Tür, Dilek Z. Hakkani-Tür, S. Bangalore, G. Riccardi, and M. Gilbert. The at&t spoken language understanding system. *IEEE Transactions on Audio, Speech, and Language Processing*, 14:213–222, 2006.

Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. Semantic parsing for task oriented dialog using hierarchical representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2787–2792, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy, July 2019. Association for Computational Linguistics.

Dilek Hakkani-Tür, Gökhan Tür, Asli Celikyilmaz, Yun-Nung Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang. Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. In *Interspeech*, pages 715–719, 2016.

Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 583–592, Online, July 2020. Association for Computational Linguistics.

Victor Petrén Bach Hansen and Anders Søgaard. What do you mean 'why?': Resolving sluices in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7887–7894, 2020.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42 (1-3):335–346, 1990.

Robert X. D. Hawkins, H. Gweon, and Noah D. Goodman. The division of labor in communication: Speakers help listeners account for asymmetries in visual perspective. *Cognitive science*, 45 3:e12926, 2021.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1766–1776. Association for Computational Linguistics, 2017.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

Matthew Henderson. Machine learning for dialog state tracking: A review. In *The First International Workshop on Machine Learning in Spoken Language Processing*, 2015.

Matthew Henderson, Blaise Thomson, and Jason D. Williams. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272, Philadelphia, PA, U.S.A., June 2014a. Association for Computational Linguistics.

Matthew Henderson, Blaise Thomson, and Steve Young. Word-based dialog state tracking with recurrent neural networks. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 292–299, Philadelphia, PA, U.S.A., June 2014b. Association for Computational Linguistics.

Jos Hernndez-Orallo. *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. Cambridge University Press, New York, NY, USA, 1st edition, 2017. ISBN 1107153018, 9781107153011.

Annette Herskovits. *Language and spatial cognition*. Cambridge university press, 1987.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019.

Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *CoRR*, abs/1508.01991, 2015.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019.

Taichi Iki and Akiko Aizawa. Language-conditioned feature pyramids for visual selection tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4687–4697, 2020.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. Meetup! a corpus of joint activity dialogues in a visual environment. *arXiv preprint arXiv:1907.05084*, 2019.

Youngsoo Jang, Jongmin Lee, and Kee-Eung Kim. Bayes-adaptive monte-carlo planning and learning for goal-oriented dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7994–8001, 2020.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

H. Kamp. A theory of truth and semantic representation, 277-322, jag groenendijk, tmv janssen and mbj stokhof, eds. In Jeroen Groenendijk, editor, *Formal Methods in the Study of Language*. U of Amsterdam, 1981.

Hans Kamp and Uwe Reyle. *From Discourse to Logic. Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer, Dordrecht, 1993.

Gi-Cheon Kang, Jaeseo Lim, and Byoung-Tak Zhang. Dual attention networks for visual reference resolution in visual dialog. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2024–2033, Hong Kong, China, November 2019. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. Learning the difference that makes a difference with counterfactually-augmented data. *International Conference on Learning Representations (ICLR)*, 2020.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798. Association for Computational Linguistics, 2014.

Boaz Keysar, Dale J Barr, Jennifer A Balin, and Jason S Brauner. Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, 11(1):32–38, 2000.

Hyounghun Kim, Hao Tan, and Mohit Bansal. Modality-balanced models for visual dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8091–8098, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Thomas Kollar, Danielle Berry, Lauren Stuart, Karolina Owczarzak, Tagyoung Chung, Lambert Mathias, Michael Kayser, Bradford Snow, and Spyros Matsoukas. The Alexa meaning representation language. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 177–184, New Orleans - Louisiana, June 2018. Association for Computational Linguistics.

Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens. Spatial role labeling: Task definition and annotation scheme. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010. European Language Resources Association (ELRA).

Satwik Kottur, José Moura, Stefan Lee, and Dhruv Batra. Natural language does not emerge 'naturally' in multi-agent dialog. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2962–2967, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Satwik Kottur, Jose M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Visual coreference resolution in visual dialog using neural module networks. In *The European Conference on Computer Vision (ECCV)*, September 2018.

Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. CLEVR-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 582–595, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*, 2017a.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017b.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

B. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the International Conference on Machine Learning*, 2018.

George Lakoff. *Women, Fire and Dangerous Things: What Categories Reveal About the Mind*. University of Chicago Press, 1987.

Barbara Landau. Update on "what" and "where" in spatial language: A new division of labor for spatial terms. *Cognitive science*, 41:321–350, 2017.

Barbara Landau and Ray Jackendoff. "what" and "where" in spatial language and spatial cognition. *Behavioral and brain sciences*, 16(2):217–238, 1993.

Staffan Larsson. Grounding as a side-effect of grounding. *Topics in cognitive science*, 10(2):389–408, 2018.

Alex Lascarides and Nicholas Asher. Agreement, disputes and commitments in dialogue. *Journal of semantics*, 26(2):109–158, 2009.

Alex Lascarides and Matthew Stone. A formal semantic analysis of gesture. *Journal of Semantics*, 26(4):393–449, 2009.

Hung Le, Doyen Sahoo, Nancy Chen, and Steven C.H. Hoi. BiST: Bi-directional spatio-temporal reasoning for video-grounded dialogues. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1846–1859, 2020.

S Lee, H Schulz, A Atkinson, J Gao, K Suleman, L El Asri, M Adada, M Huang, S Sharma, W Tay, et al. Multi-domain task-completion dialog challenge. *Dialog system technology challenges*, 8:9, 2019.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium, 2018. Association for Computational Linguistics.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. TVQA+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225, Online, July 2020. Association for Computational Linguistics.

David Lewis. Convention cambridge. *Mass.: Harvard UP*, 1969.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

Yu Li, Kun Qian, Weiyan Shi, and Zhou Yu. End-to-end trainable non-collaborative dialog system. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8293–8302. AAAI Press, 2020.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016.

Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132. Association for Computational Linguistics, 2016.

Edward Loper and Steven Bird. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. Towards an automatic Turing test: Learning to

evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada, July 2017. Association for Computational Linguistics.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019b.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2): 313–330, 1993.

Daniela Massiceti, Puneet K Dokania, N Siddharth, and Philip HS Torr. Visual dialogue without vision or dialogue. *arXiv preprint arXiv:1812.06417*, 2018.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics.

Bill McDowell and Noah Goodman. Learning from omission. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 619–628, Florence, Italy, July 2019. Association for Computational Linguistics.

Grégoire Mesnil, Xiaodong He, Li Deng, and Yoshua Bengio. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *Interspeech*, pages 3771–3775, 2013.

Will Monroe, Robert X. D. Hawkins, Noah D. Goodman, and Christopher Potts. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338, 2017.

Seungwhan Moon, Satwik Kottur, Paul Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. Situated and interactive multimodal conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1103–1121, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. *arXiv preprint arXiv:1912.02379*, 2019.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Yukiko Nakano, Gabe Reinstein, Tom Stocky, and Justine Cassell. Towards a model of face-to-face grounding. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 553–561, Sapporo, Japan, July 2003. Association for Computational Linguistics.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. Collaborative dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415, Florence, Italy, July 2019. Association for Computational Linguistics.

Vincent Ng. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328, Melbourne, Australia, 2018. Association for Computational Linguistics.

Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. Recursive visual attention in visual dialog. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252. Association for Computational Linguistics, 2017.

Christopher Olah. Understanding LSTM networks, 2015. URL `http://colah.github.io/posts/2015-08-Understanding-LSTMs/`. Last accessed on June 4, 2021.

Wei Pang and Xiaojie Wang. Visual dialogue state tracking for question generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11831–11838, 2020.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

Carita Paradis. Configurations, construals and change: expressions of DEGREE. *English Language and Linguistics*, 12(2):317–343, 2008.

Ramakanth Pasunuru and Mohit Bansal. Game-based video-context dialogue. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 125–136, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

Baolin Peng, Xiujun Li, Lihong Li, Jianfeng Gao, Asli Celikyilmaz, Sungjin Lee, and Kam-Fai Wong. Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2231–2240, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Miriam R. L. Petruck and Michael J. Ellsworth. Representing spatial relations in FrameNet. In *Proceedings of the First International Workshop on Spatial Language Understanding*, pages 41–45, New Orleans, June 2018. Association for Computational Linguistics.

Martin J Pickering and Simon Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190, 2004.

Georgiy Platonov and Lenhart Schubert. Computational models for spatial prepositions. In *Proceedings of the First International Workshop on Spatial Language Understanding*, pages 21–30, 2018.

Massimo Poesio and Hannes Rieser. Completions, coordination, and alignment in dialogue. *Dialogue & Discourse*, 1(1), 2010.

Massimo Poesio and David R Traum. Conversational actions and discourse situations. *Computational intelligence*, 13(3):309–347, 1997.

Massimo Poesio, Roland Stuckardt, and Yannick Versley. *Anaphora resolution*. Springer, 2016.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1778–1789, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Christopher Potts. Goal-driven answers in the cards dialogue corpus. In *Proceedings of the 30th west coast conference on formal linguistics*, pages 1–20. Cascadilla Proceedings Project, 2012.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27. Association for Computational Linguistics, 2011.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3: 28–34, 2003.

James Pustejovsky, Jessica L Moszkowicz, and Marc Verhagen. Iso-space: The annotation of spatial information in language. In *Proceedings of the Sixth Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, volume 6, pages 1–9, 2011a.

James Pustejovsky, Jessica L Moszkowicz, and Marc Verhagen. Using iso-space for annotating spatial information. In *Proceedings of the International Conference on Spatial Information Theory*, 2011b.

James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. Semeval-2015 task 8: Spaceeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (semeval 2015)*, pages 884–894. ACL, 2015.

Jun Quan, Deyi Xiong, Bonnie Webber, and Changjian Hu. GECOR: An end-to-end generative ellipsis and co-reference resolution model for task-oriented dialogue. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4547–4557, Hong Kong, China, November 2019. Association for Computational Linguistics.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*, 2018.

Arnau Ramisa, Josiah Wang, Ying Lu, Emmanuel Dellandrea, Francesc Moreno-Noguer, and Robert Gaizauskas. Combining geometric, textual and visual features for predicting prepositions in image descriptions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 214–220, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696, 2020.

Marta Recasens, M. Antònia Martí, and Constantin Orasan. Annotating near-identity from coreference disagreements. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 165–172, Istanbul, Turkey, 2012. European Languages Resources Association (ELRA).

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online, April 2021. Association for Computational Linguistics.

Harvey Sacks. On the preferences for agreement and contiguity in sequences in conversation. *Talk and social organisation*, pages 54–69, 1987.

Harvey Sacks, E. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696–735, 1974.

Arka Sadhu, Kan Chen, and Ram Nevatia. Video object grounding using semantic roles in language description. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Ananya B Sai, Mithun Das Gupta, Mitesh M Khapra, and Mukundhan Srinivasan. Re-evaluating adem: A deeper look at scoring dialogue responses. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6220–6227, 2019.

Chinnadhurai Sankar, Sandeep Subramanian, Chris Pal, Sarath Chandar, and Yoshua Bengio. Do neural dialog systems use the conversation history effectively? an empirical study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 32–37, Florence, Italy, July 2019. Association for Computational Linguistics.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Tim Lillicrap. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pages 4967–4976, 2017.

Emanuel A Schegloff, Gail Jefferson, and Harvey Sacks. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382, 1977.

T.C. Schelling. *The Strategy of Conflict*. Harvard University Press, 1960.

David Schlangen. Grounded agreement games: Emphasizing conversational grounding in visual dialogue settings. *arXiv preprint arXiv:1908.11279*, 2019.

John R Searle. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press, 1969.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 3776–3783. AAAI Press, 2016.

Igor Shalyminov, Alessandro Sordoni, Adam Atkinson, and Hannes Schulz. Fast domain adaptation for goal-oriented dialogue using a hybrid generative-retrieval transformer. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2020.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2556–2565, 2018.

Ravi Shekhar, Aashish Venkatesh, Tim Baumgärtner, Elia Bruni, Barbara Plank, Raffaella Bernardi, and Raquel Fernández. Beyond task success: A closer look at jointly learning to see, ask, and GuessWhat. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2578–2587, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Yangyang Shi, Kaisheng Yao, Le Tian, and Daxin Jiang. Deep LSTM based feature mapping for query classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1501–1511, San Diego, California, June 2016. Association for Computational Linguistics.

Todd Shore, Theofronia Androulakaki, and Gabriel Skantze. KTH tangrams: A dataset for research on alignment and conceptual pacts in task-oriented dialogue. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

Pushkar Shukla, Carlos Elmadjian, Richika Sharan, Vivek Kulkarni, Matthew Turk, and William Yang Wang. What should I ask? using conversationally informative rewards for goal-oriented visual dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6442–6451, Florence, Italy, July 2019. Association for Computational Linguistics.

Joao Silva, Luísa Coheur, Ana Cristina Mendes, and Andreas Wichert. From symbolic to sub-symbolic information in question classification. *Artificial Intelligence Review*, 35(2):137–154, 2011.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online, July 2020. Association for Computational Linguistics.

Robert C Stalnaker. *Assertion*. Wiley Online Library, 1978.

L. Steels and Tony Belpaeme. coordinating perceptually grounded categories through language: a case study for colour. *Behavioral and Brain Sciences*, 28:469–489, 2005.

Luc Steels. The synthetic modeling of language origins. *Evolution of communication*, 1 (1):1–34, 1997.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics, 2012.

Amanda Stent. Rhetorical structure in dialog. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 247–252, Mitzpe Ramon, Israel, June 2000. Association for Computational Linguistics.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374, 2000.

Saku Sugawara, Hikaru Yokono, and Akiko Aizawa. Prerequisite skills for reading comprehension: Multi-perspective analysis of mctest datasets and systems. In *AAAI*, pages 3089–3096, 2017.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. What makes reading comprehension questions easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4219, Brussels, Belgium, 2018. Association for Computational Linguistics.

Saku Sugawara, Pontus Stenetorp, and Akiko Aizawa. Benchmarking machine reading comprehension: A psychological perspective. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1592–1612, Online, April 2021. Association for Computational Linguistics.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 217–223, 2017.

Alane Suhr, Claudia Yan, Jack Schluger, Stanley Yu, Hadi Khader, Marwa Mouallem, Iris Zhang, and Yoav Artzi. Executing instructions in situated collaborative interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2119–2130, Hong Kong, China, November 2019. Association for Computational Linguistics.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Ryuichi Takanobu, Hanlin Zhu, and Minlie Huang. Guided dialog policy learning: Reward estimation for multi-domain task-oriented dialog. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 100–110, Hong Kong, China, November 2019. Association for Computational Linguistics.

Ryuichi Takanobu, Qi Zhu, Jinchao Li, Baolin Peng, Jianfeng Gao, and Minlie Huang. Is your goal-oriented dialog model performing really well? empirical analysis of system-wise evaluation. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 297–310, 1st virtual meeting, July 2020. Association for Computational Linguistics.

Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, and Raquel Fernández. Refer, reuse, reduce: Generating subsequent references in visual and conversational contexts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4350–4368, 2020.

Leonard Talmy. How language structures space. In *Spatial orientation*, pages 225–282. Springer, 1983.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *Conference on Robot Learning (CoRL)*, 2019.

Takenobu Tokunaga, Ryu Iida, Asuka Terai, and Naoko Kuriyama. The REX corpora: A collection of multimodal corpora of referring expressions in collaborative problem solving dialogues. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 422–429, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

Michael Tomasello. *Constructing a language*. Harvard university press, 2009.

David R Traum. A computational theory of grounding in natural language conversation. Technical report, ROCHESTER UNIV NY DEPT OF COMPUTER SCIENCE, 1994.

A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950. ISSN 00264423.

Takuma Udagawa and Akiko Aizawa. A natural language corpus of common grounding under continuous and partially-observable context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7120–7127, 2019.

Takuma Udagawa and Akiko Aizawa. An annotated corpus of reference resolution for interpreting common grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9081–9089, 2020.

Takuma Udagawa and Akiko Aizawa. Maintaining common ground in dynamic environments. *Transactions of the Association for Computational Linguistics*, 9, 2021.

Takuma Udagawa, Takato Yamazaki, and Akiko Aizawa. A linguistic analysis of visually grounded dialogues based on spatial expressions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 750–765, Online, November 2020. Association for Computational Linguistics.

Morgan Ulinski, Bob Coyne, and Julia Hirschberg. SpatialNet: A declarative resource for spatial relations. In *Proceedings of the Combined Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP)*, pages 61–70, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. Learning to speak and act in a fantasy text adventure game. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 673–683, Hong Kong, China, November 2019. Association for Computational Linguistics.

Hans Van Ditmarsch, Wiebe van Der Hoek, and Barteld Kooi. *Dynamic epistemic logic*, volume 337. Springer Science & Business Media, 2007.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. PARADISE: A framework for evaluating spoken dialogue agents. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280, Madrid, Spain, July 1997. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018a. Association for Computational Linguistics.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018b.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy, July 2019. Association for Computational Linguistics.

Zhongdao Wang, Liang Zheng, Yixuan Liu, and Shengjin Wang. Towards real-time multi-object tracking. In *European Conference on Computer Vision*, 2020.

Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain, April 2017. Association for Computational Linguistics.

Jason Williams, Antoine Raux, and Matthew Henderson. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33, 2016.

Jason D. Williams, Kavosh Asadi, and Geoffrey Zweig. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 665–677, Vancouver, Canada, July 2017. Association for Computational Linguistics.

Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, and Anton van den Hengel. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*, pages 2048–2057, 2015.

Rui Yan, Yiping Song, and Hua Wu. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39th*

*International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 55–64, 2016.

Kaisheng Yao, Geoffrey Zweig, Mei-Yuh Hwang, Yangyang Shi, and Dong Yu. Recurrent neural networks for language understanding. In *Interspeech*, pages 2524–2528, 2013.

Denis Yarats and Mike Lewis. Hierarchical text generation and planning for strategic dialogue. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 5587–5595, 2018.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, 2020.

Koichiro Yoshino, Shinsuke Mori, and Tatsuya Kawahara. Spoken dialogue system based on information extraction using similarity of predicate argument structures. In *Proceedings of the SIGDIAL 2011 Conference*, pages 59–66. Association for Computational Linguistics, 2011.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179, 2013.

Xintong Yu, Hongming Zhang, Yangqiu Song, Yan Song, and Changshui Zhang. What you see is what you get: Visual pronoun coreference resolution in dialogues. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5123–5132, Hong Kong, China, November 2019a. Association for Computational Linguistics.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, pages 9127–9134, 2019b.

Sina Zarrieß, Julian Hough, Casey Kennington, Ramesh Manuvinakurike, David DeVault, Raquel Fernández, and David Schlangen. PentoRef: A corpus of spoken references in task-oriented dialogues. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 125–131, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6720–6731, 2019.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online, July 2020. Association for Computational Linguistics.

Zilong Zheng, Wenguan Wang, Siyuan Qi, and Song-Chun Zhu. Reasoning visual dialogs with structural and partial observations. In *Computer Vision and Pattern Recognition (CVPR), 2019 IEEE Conference on*, 2019.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93, 2020.

Luowei Zhou, Nathan Louis, and Jason J Corso. Weakly-supervised video object grounding from text by loss weighting and object interaction. In *British Machine Vision Conference*, 2018.

Rolf A Zwaan and Gabriel A Radvansky. Situation models in language comprehension and memory. *Psychological bulletin*, 123(2):162, 1998.