Meet Up! A Corpus of Joint Activity Dialogues in a Visual Environment

Nikolai Ilinykh
Dialogue Systems Group
Bielefeld University
nikolai.ilinykh@uni-bielefeld.de

Sina Zarrieß *
Digital Humanities
University of Jena
sina.zarriess@uni-jena.de

David Schlangen *
Computational Linguistics
University of Potsdam
david.schlangen@uni-potsdam.de

Abstract

Building computer systems that can converse about their visual environment is one of the oldest concerns of research in Artificial Intelligence and Computational Linguistics (see, for example, Winograd's 1972 SHRDLU system). Only recently, however, have methods from computer vision and natural language processing become powerful enough to make this vision seem more attainable. Pushed especially by developments in computer vision, many data sets and collection environments have recently been published that bring together verbal interaction and visual processing. Here, we argue that these datasets tend to oversimplify the dialogue part, and we propose a task—MeetUp!—that requires both visual and conversational grounding, and that makes stronger demands on representations of the discourse. MeetUp! is a two-player coordination game where players move in a visual environment, with the objective of finding each other. To do so, they must talk about what they see, and achieve mutual understanding. We describe a data collection and show that the resulting dialogues indeed exhibit the dialogue phenomena of interest, while also challenging the language & vision aspect.

1 Introduction

In recent years, there has been an explosion of interest in language & vision in the NLP community, leading to systems and models able to ground the meaning of words and sentences in visual representations of their corresponding referents, e.g. work in object recognition (Szegedy et al., 2015), image captioning (Fang et al., 2015; Devlin et al., 2015; Chen and Lawrence Zitnick, 2015; Vinyals et al., 2015; Bernardi et al., 2016), referring expression resolution and generation (Kazemzadeh et al., 2014; Mao et al., 2015; Yu et al., 2016; Schlangen et al., 2016), multi-modal distributional semantics (Kiela and Bottou, 2014; Silberer and Lapata, 2014; Lazaridou et al., 2015), and many others.

While these approaches focus entirely on visual grounding in a static setup, a range of recent initiatives have extended exisiting data sets and models to more interactive settings. Here, speakers do not only describe a single image or object in an isolated utterance, but engage in some type of multi-turn interaction to solve a given task (Das et al., 2017b; De Vries et al., 2017). In theory, these data sets should allow for more dynamic approaches to *grounding* in natural language interaction, where words or phrases do not simply have a static multi-modal meaning (as in existing models for distributional semantics, for instance), but, instead, where the meaning of an utterance is *negotiated* and *established* during interaction. Thus, ideally, these data sets should lead to models that combine visual grounding in the sense of Harnard (1990) and conversational grounding in the sense of Clark et al. (1991).

In practice, however, it turns out to be surprisingly difficult to come up with data collection setups that lead to interesting studies of both these aspects of grounding. Existing tasks still adopt a very rigid interaction protocol, where e.g. an asymmetric interaction between a question asker and a question answerer produces uniform sequences of question-answer pairs (as in the "Visual Dialogue" setting of Das et al. (2017b) for instance). Here, it is impossible to model e.g. turn-taking, clarification, collaborative utterance construction, which are typical phenomena of conversational grounding in interaction (Clark, 1996b). Others tasks follow the traditional idea of the *reference game* (Rosenberg and Cohen, 1964; Clark and Wilkes-Gibbs, 1986) in some way, but try to set up the game such that the referent can only be

*Work done while at Bielefeld University.

established in a sequence of turns (e.g. De Vries et al., 2017). While this approach leads to goal-oriented dialogue, the goal is still directly related to reference and visual grounding. However, realistic, every-day communication between human speakers rarely centers entirely around establishing reference. It has been argued in the literature that reference production radically changes if it is the primary goal of an interactive game, rather than embedded in a dialogue that tries to achieve a more high-level communicative goal (Stent, 2011).

Another strand of recent work extends the environments about which the language can talk to (simulated) 3D environments (Savva et al. (2019, 2017); see Byron et al. (2007) for an early precursor). On the language side, however, the tasks that have been proposed in these environments allow only limited interactivity (navigation, e.g. Anderson et al. (2018); Ma et al. (2019); question answering, Das et al. (2017a)).

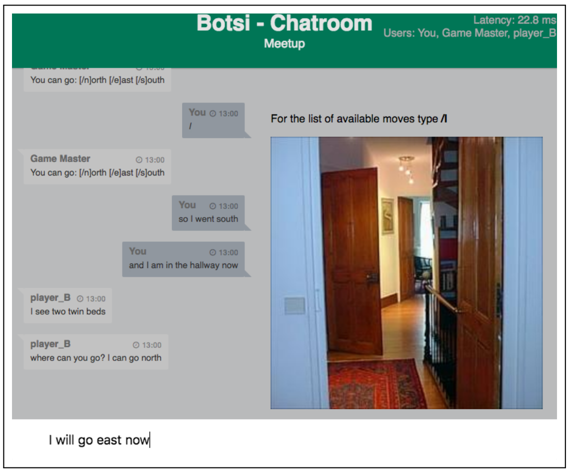

Figure 1: The game interface

What is lacking in these tasks is a real sense of the interaction being a *joint task* for which both participants are equally responsible, and, phrased more technically, any need for the participants to jointly attempt to track the dialogue state. In this paper, we propose a new task, MeetUp!, for visually grounded interaction, which is aimed at collecting conversations about and within a visual world, in a collaborative setting. (Figure 1 gives a view of the game interface and an excerpt of an ongoing interaction.)

Our setup extends recent efforts along three main dimensions: 1) the task's main goal can be defined independently of reference, in high-level communicative terms (namely "try to meet up in an unknown environment"), 2) the task is symmetric and does not need a rigid interaction protocol (there is no instruction giver/follower), 3) the requirement to *agree* on the game state (see below) ensures that the task is a true *joint activity* (Clark, 1996a), which in turn brings out opportunity for *meta-semantic* interaction and negotiation about perceptual classifications ("there is a mirror" – "hm, could it be a picture?". This is an important phenomenon absent from all major current language & vision datasets.

This brings our dataset closer to those of unrestricted natural situated dialogue, e.g. (Anderson et al., 1991; Fernández and Schlangen, 2007; Tokunaga et al., 2012; Zarrieß et al., 2016), while still affording us some control over the expected range of phenomena, following our design goal of creating a challenging, but not too challenging modelling resource. The crowd-sourced nature of the collection also allows us to create a resource that is an order of magnitude larger than those just mentioned.[1]

We present our data collection of over 400 dialogues in this domain, providing an overview of the characteristics and an analysis of some occuring phenomena. Results indicate that the task leads to rich, natural and varied dialogue where speakers use a range of strategies to achieve communicative grounding. The data is available from `https://github.com/clp-research/meetup`.

2 The Meet Up Game

MeetUp! is a two-player coordination game. In the discrete version described here, it is played on a gameboard that can be formalised as a connected subgraph of a two-dimensional grid graph.[2] See Figure 2 for an example.

Players are located at vertices in the graph,

[1] Haber et al. (2019) present a concurrently collected dataset that followed very similar aims (and is even larger); their setting however does not include any navigational aspects and concentrates on reaching agreement of whether images are shared between the participants or not.

[2] The game could also be realised in an environment that allows for continuous movement and possibly interaction with objects, for example as provided by the simulators discussed above. This would complicate the navigation and visual grounding aspects (bringing those more in line with the "vision-and-language navigation task"; (e.g. Anderson et al., 2018; Ma et al., 2019)), but not the coordination aspect. As our focus for now is on the latter, we begin with the discrete variant.

which we call “rooms”. Players never see a representation of the whole gameboard, they only see their current room (as an image). They also do not see each other’s location. The images representing rooms are of different types; here, different types of real-world scenes, such as “bathroom”, “garage”, etc., taken from the ADE20k corpus collected by Zhou et al. (2017). Players can move from room to room, if there is a connecting edge on the gameboard. On entering a room, the player is (privately) informed about the available exit directions as cardinal directions, e.g. “north”, “south”, etc., and (privately) shown the image that represents the room. Players move by issuing commands to the game; these are not shown to the other player.

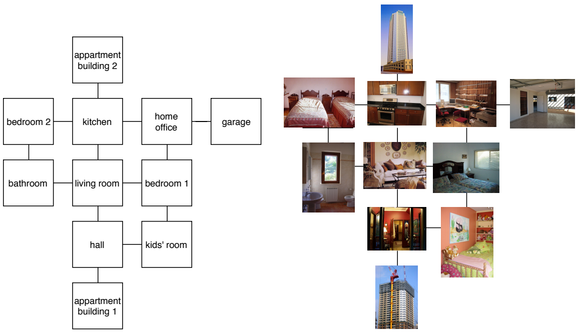

Figure 2: An abstract layout with room types (left), and a full gameboard with assigned images (right).

The goal of the players is to be in the same location, which means they also have to be aware of that fact. In the variant explored here, the goal is constrained in advance in that the meetup room has to be of a certain type previously announced to the players; e.g., a kitchen. The players can communicate via text messages. As they do not see each other’s location, they have to describe the images they see to ascertain whether or not they are currently in the same room, and move to a different room if they decide that they aren’t. If they have reached the conclusion that they are, they can decide to end the game, which they do via a special command. If they are then indeed in the the same room, and it is of the target type, the game is counted as a success, of which they are informed. The gameboard can be arranged such that there is type-level ambiguity; for example, there may be more than one room of type “bedroom” (as in Figure 2).

The game as implemented does not impose strict turn taking on the players; however, messages are only shown to the other player once they are sent via pressing the return key, as is usual in chat tools. There is thus no possibility for perceivably overlapping actions, but it may happen that both players have been typing at the same time and the message that is received second is not a response to the first.

To make this more concrete, and to explain our expectations with respect to phenomena and required capabilities, we show a realistic, but compressed and constructed example of an interaction in this domain in the following. We will discuss attested examples from our data collection further below.

(1)
a. Game Master: You have to meet in a room of type *utility room*.
b. A: Hi. I’m in a bedroom with pink walls.
c. B: I seem to be in a kitchen.
d. A: I’ll go look for a utility room.
e. A (privately): *north*
f. A (privately): *west*
g. B (privately): *east*
h. A: Found a room with a washing machine. Is that a utility room?
i. B: Was wondering as well. Probably that’s what it is.
j. B: I’m in the pink bedroom now. I’ll come to you.
k. B (privately): *north*
l. B (privately): *west*
m. B: Poster above washing machine?
n. A: Mine has a mirror on the wall.
o. B: yeah, could be mirror. Plastic chair?
p. A: And laundry basket.
q. A: *done*
r. B: Same
s. B: *done*

In (1-a), the Game Master (realised as a software bot in the chat software) gives the type constraint for the meetup room, which sets up a **classification task** for the players, namely to identify rooms of this type. (1-b) and (1-c) illustrate a common strategy (as we will see below), which is to start the interaction by providing state information that potentially synchronises the mutual representations. This is done through the production of **high-level descriptions of the current room**; for which the agents must be capable of providing *scene categorisations*. (1-d) and (1-j) show, among other things, the **coordination of strategy**, by announcing plans for action. In (1-e) – (1-g), private navigation actions are performed, which here are both **epistemic actions** (changing the environment to change perceptual state) as well as **pragmatic actions** (task level actions that potentially advance towards the goal), in the sense of Kirsh and Maglio (1994). (1-h) and (1-i), where the classification decision itself and its basis is discussed ("what is a utility room?"); and (1-m)–(1-o), where a classification decision is revised (*poster* to *mirror*), illustrate the potential for **meta-semantic interaction**. This is an important type of dialogue move (Schlangen, 2016), which is entirely absent from most other language and vision datasets and hence outside of the scope of models trained on them. (1-j), also illustrates the need for **discourse memory**, through the co-reference to the earlier mentioned room where A was at the start. Finally, (1-p) as reply to (1-o) shows how in conversational language, **dialogue acts** can be **performed indirectly**.

As we have illustrated with this constructed example, the expectation is that this domain challenges a wide range of capabilities; capabilities which so far have been captured separately (e.g., visual question answering, scene categorisation, navigation based on natural language commands, discourse co-reference), or not at all (discussion and revision of categorisation decisions). We will see in the next section whether this is borne out by the data.

3 Data Collection

To test our assumptions, and to later derive models for these phenomena, we collected a larger number of dialogues in this domain (430, to be precise). We realised the MeetUp game within the *slurk* chat-tool (Schlangen et al., 2018), deployed via the Amazon Mechanical Turk platform.

We constructed maps for the game in three steps. First, we create a *graph* through a random walk over a grid graph, constrained to creating 10 nodes. The nodes are then assigned room types, to form what we call a *layout*. We identified 48 categories from the ADE20k corpus that we deemed plausible to appear in a residential house setting, from which we designated 20 categories as possible (easy to name) target types and the remaining 28 as distractor types. Additionally, we identified 24 plausible outdoor scene types, from which we sampled for the leaf nodes. The full set is given in the Appendix. We designate one type per layout to be the target type; this type will be assigned to 4 nodes in the graph, to achieve type ambiguity and potentially trigger clarification phases. We then sample actual images from the appropriate ADE20k categories, to create the *gameboards*. In a final step, we randomly draw separate starting positions for the players, such that both of the players start in rooms not of the target type. For each run of the game, we randomly create a new gameboard following this recipe.

We deployed the game as a web application, enlisting workers via the Mechanical Turk platform. After reading a short description of the game (similar to that at the beginning of Section 2, but explaining the interface in more detail), workers who accepted the task were transferred to a waiting area in our chat tool. If no other worker appeared within a set amount of time, they were dismissed (and payed for their waiting time). Otherwise, the pair of users was moved to another room in the chat tool and the game begun. Player were payed an amount of \$0.15 per minute (for a maximum of 5 minutes per game), with a bonus of \$0.10 for successfully finishing the game (as was explained from the start in the instruction, to provide an additional incentive).[3]

4 Results

4.1 Descriptive Statistics

Over a period of 4 weeks, we collected 547 plays of the game. Of these, 117 (21%) had to be discarded because one player left prematurely or technical problems occurred, which left us with 430 completed dialogues. Of these, 87% ended successfully (players indeed ending up in the same room, of the correct type), 10% ended with the players being in different rooms of the correct type; the remaining 3% ended with at least one player not even being in a room of the target type. Overall, we spent around \$700 on the data collection.

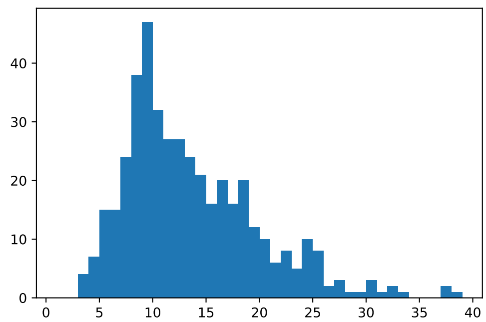

Figure 3: Histogram of number of turns per dialogue

[3]By the time of the conference, we will publish the code required to run this environment, as well as the data that we collected.

The average length of a dialogue was 13.2 turns (66.9 tokens), taking 165 seconds to produce. (The distribution of lengths is shown in Figure 3.) Altogether, we collected 5,695 turns, of an average length of 5.1 tokens. Over all dialogues, 2,983 word form types were introduced, leading to a type/token ratio of 0.10. The overlap of the vocabularies of the two players (intersection over union) ranged from none to 0.5, with a mean of 0.11.

On average, in each dialogue 28.3 navigation actions were performed. (Resulting in a MOVE/SAY ratio of a little over 2 to 1). The median time spent in a room was 12.2 secs. On average, each player visited 5.9 rooms without saying anything; when a player said something while in a room, they produced on average 3.5 turns. It hence seems that, as expected, players moved through some rooms without commenting on them, while spending more time in others.

We calculated the contribution ratio between the more talkative player and the less talkative one in each dialogue, which came out as 2.4 in terms of tokens, and 1.7 in terms of turns. This indicates that there was a tendency for one of the players to take a more active role. To provide a comparison, we calculated the same for the (role-asymmetric) MapTask dialogues (Anderson et al., 1991),[4] finding a 2.8 token ratio and a 1.3 turn ratio.

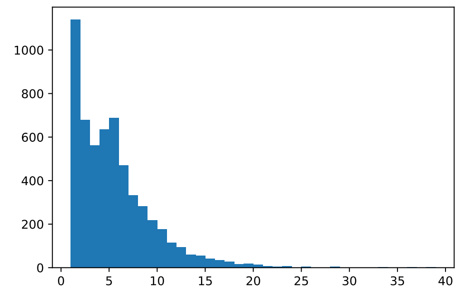

Figure 4: Histogram of number of tokens per turn

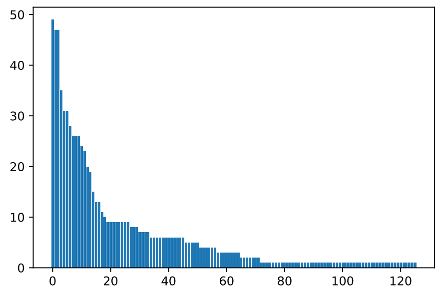

Figure 5: Number of Games Played, by Worker

Crosstalk occurs: On average, there are 1.4 instances of one turn coming within two seconds or less than the previous one (which we arbitrarily set as the threshold for when a turn is likely not to be a reaction to the previous one, but rather has been concurrently prepared). The mean pause duration between turns of different speakers is 11.2 secs – with a high standard deviation of 9.46, however. This is due to the structure of the dialogues with phases of intense communicative activity, when a matching decision is made, and phases of individual silent navigation. If we only take transition times within the first 3 quartiles, the average transition time is 5.04 secs.

As Figure 4 indicates, most turns are rather short, but there is a substantial amount of turns that contain 4 or more tokens.

Figure 5 shows a frequency distribution of number of games played, by crowdworker. Overall, we had 126 distinct participants (as indicated by AMT ID). Our most prolific worker participated in 49 games, and the majority of workers played in more than one game. In only 22 games, two novices played with each other. In 81 games, there was one novice, and in 305 games, both players had played before. (For a few games we could not reconstruct the workerIDs for technical reasons, so this does not sum up to 430.)

4.2 Examples

Figure 6 shows a full interaction from the corpus. The public actions are represented in bold font, private actions are marked with “privately”, and responses by the Game Master are shown in italics. This example has many of the features illustrated with the constructed example (1) shown earlier. In lines 20 and 22, the players begin the game by providing high-level categorisations of their current positions, in effect synchronising their mutual game representations. Lines 22 and 23 then show coordination of game playing strategy. After separately moving around, individually solving the categorisation task–by moving through rooms that are not of the right type—the players resume interaction in lines 50ff. (with 50/51 showing crosstalk). Line 54 provides a justification for the negative conclusion from line 53, by providing information that contradicts l. 52. After more coordination of strategy in l.s 55 & 56, player B explicitly queries for more information. In line 75, player A justifies their positive verdict by confirming details previously given by B, extending it with

[4] Using the transcripts provided at `http://groups.inf.ed.ac.uk/maptask/maptasknxt.html`.

	time	main
10	00:03	*GM (to A): url: /b/bathroom/ADE-train-00002739.jpg*
12	00:03	*GM (to A): You can go: [/s]outh [/n]orth [/w]est*
14	00:03	*GM (to A): You have to meet in: b/basement*
15	00:03	*GM (to B): url: /a/apartment-building/outdoor/ADE-train-00001414.jpg*
17	00:04	*GM (to B): You can go: [/s]outh*
19	00:04	*GM (to B): You have to meet in: b/basement*
20	00:18	**A: i am in the bathroom**
21	00:29	**A: where are you at?**
22	00:33	**B: I'm outside. Let's move around a bit.**
23	00:38	**A: k**
24	00:40	B (privately): n
25	00:41	*GM (to B): Nothing happened.You can go: [/s]outh*
26	00:43	A (privately): s
27	00:44	*GM (to A): url: /b/bedroom/ADE-train-00003553.jpg*
29	00:44	*GM (to A): You can go: [/s]outh [/n]orth [/w]est*
30	00:45	B (privately): s
31	00:45	*GM (to B): url: /b/basement/ADE-train-00002497.jpg*
33	00:45	*GM (to B): You can go: [/s]outh [/n]orth [/e]ast*
34	00:50	A (privately): s
35	00:50	*GM (to A): url: /c/corridor/ADE-train-00006285.jpg*
37	00:51	*GM (to A): You can go: [/n]orth [/e]ast*
38	00:52	B (privately): s
39	00:52	*GM (to B): url: /l/living-room/ADE-train-00000774.jpg*
41	00:53	*GM (to B): You can go: [/s]outh [/n]orth [/e]ast*
42	00:57	B (privately): s
43	00:57	*GM (to B): url: /b/basement/ADE-train-00002482.jpg*
45	00:57	*GM (to B): You can go: [/n]orth [/e]ast*
46	01:01	A (privately): e
47	01:01	*GM (to A): url: /b/basement/ADE-train-00002494.jpg*
49	01:01	*GM (to A): You can go: [/e]ast [/w]est*
50	01:11	**A: I am in the basement**
51	01:11	**B: I'm in a basement.**
52	01:23	**B: Mine has a white staircase**
53	01:28	**A: no**
54	01:37	**A: mine has wooden stair case**
55	01:55	**B: Okay. Should I try to move towards you?**
56	02:09	**A: Sure**
57	02:11	**B: Wooden? What else?**
58	02:16	B (privately): e
59	02:17	*GM (to B): url: /b/bedroom/ADE-train-00003553.jpg*
61	02:17	*GM (to B): You can go: [/s]outh [/n]orth [/w]est*
62	02:25	B (privately): s
63	02:25	*GM (to B): url: /c/corridor/ADE-train-00006285.jpg*
65	02:25	*GM (to B): You can go: [/n]orth [/e]ast*
66	02:27	**A: water heater and washer and dryer**
67	02:35	B (privately): s
68	02:35	*GM (to B): Nothing happened.You can go: [/n]orth [/e]ast*
69	02:40	B (privately): e
70	02:40	*GM (to B): url: /b/basement/ADE-train-00002494.jpg*
72	02:41	*GM (to B): You can go: [/e]ast [/w]est*
73	02:42	**A: a plastic chair and a screen door**
74	03:01	*GM: Attention: you are in the game for 3 minutes!*
75	03:18	**B: I'm there! I see the water heater, washer and dryer, sink, chair laying on top of the screen door**
76	03:26	B (privately): done
77	03:27	*GM: The '/done' command has been issued by you or your partner. To end the game, both players need to enter '/done'*
78	03:27	**A: yep**
79	03:32	A (privately): done
80	03:34	*GM: Well done! Both of you are indeed in the same room of type: b/basement!*

Figure 6: One Example Dialogue (mux36), with Images Overlayed

even more details. B confirms explicitly in 78, before also chosing SOLVE.

The excerpt from another dialogue in (2) shows an example of classification uncertainty being negotiated and dealt with.

(2) (Excerpt from mux39)

A: i think i am in a basement
B: i think i might be too.
A: maybe not though
A: wood panel?
A: two doors?
B: there's a tan couch, and a tan loveseat/chair. brown coffee table.

bar. tv
B: nope, different room
A: ok i am not there
B: want me to meet you, or do you want to meet me?
A: i think mine is more basement like
B: okay. i'll try to find it.

4.3 Phases and Phenomena

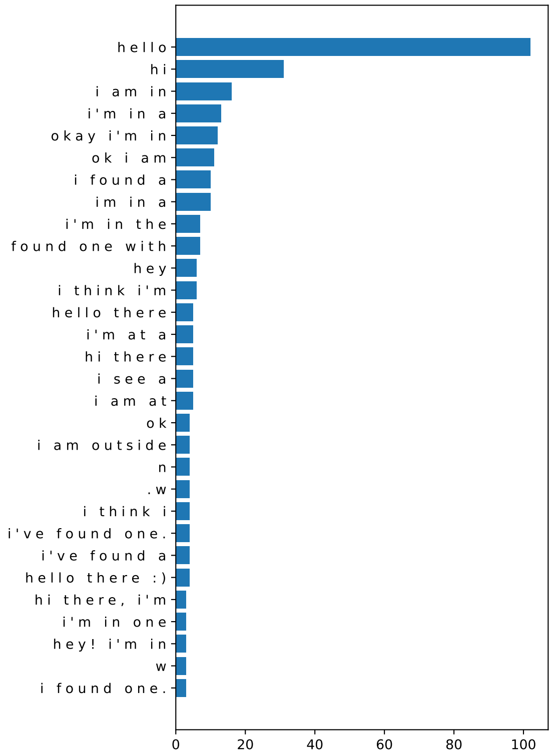

Figure 7: Prefixes of first turns

Figure 7 shows the most frequent beginnings of the very first turn in each dialogue. As this indicates, when not opening with a greeting, players naturally start by locating themselves (as in the example we showed in full). Figure 8 gives a similar view of the final turn, before the first *done* was issued. This shows that the game typically ends with an explicit mutual confirmation that the goal condition was reached, before this was indicated to the game.

What happens inbetween? Figure 9 shows the most frequent overall turn beginnings. As this illustrates, besides the frequent positive replies ("yes", "ok"; indicating a substantial involvement of VQA-like interactions), the most frequent constructions seem to locate the speaker ("I'm in a") or talk about objects ("I found a", "there is a", "is there a"). Using the presence of a question mark at

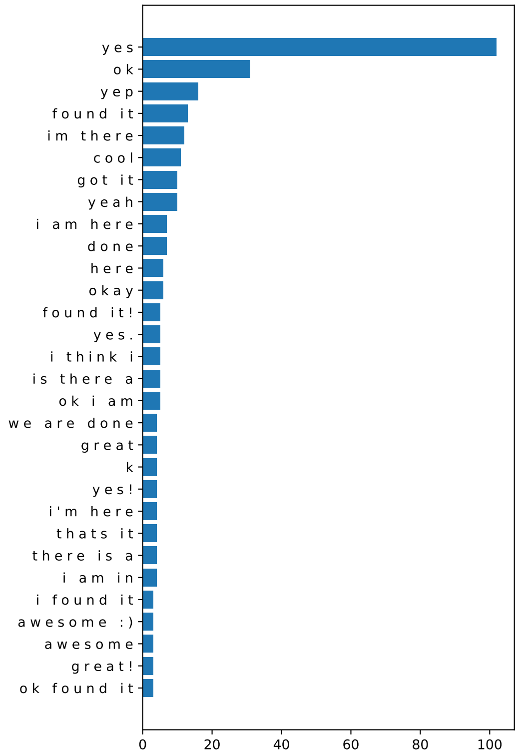

Figure 8: Prefixes of final turns (before *done*)

the end of the turn as a very rough proxy, we find 615 questions over all dialogues, which works out as 1.43 on average per dialogue. Taking only the successfull dialogues into account, the number is slightly higher, at 1.48. Figure 10 shows the beginnings of these turns.

5 Modelling the Game

The main task of an agent playing this game can be modelled in the usual way of modelling agents in dynamic environments (Sutton and Barto, 1998), that is, as computing the best possible next action, given what has been experienced so far. The questions then are what the range of possible actions is, what the agent needs to remember about its experience, and what the criteria might be for selecting the best action.

In the action space, the clearest division is between actions that are directly observable by the other player—actions of type SAY—and actions that are targeted at changing the observable game state for the agent itself: actions of type MOVE and the END action. Since we did not restrict what the players could say, there is an infinite number of SAY actions (see Côté et al. (2018) for a formalisation of such an action space).

The total game state consists of the positions of

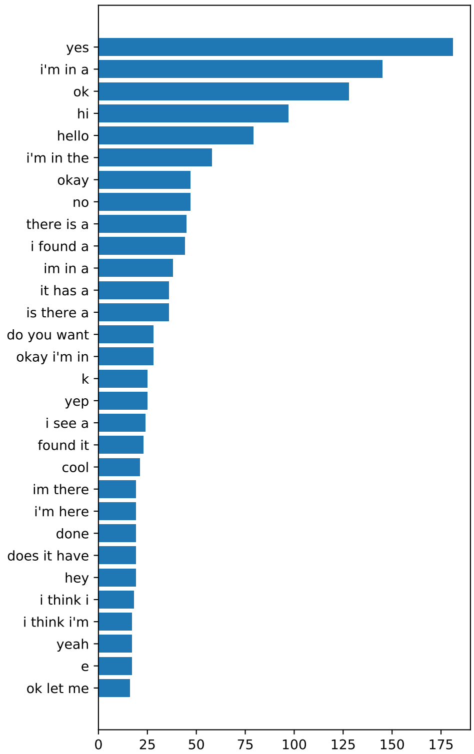

Figure 9: Most frequent turn beginnings

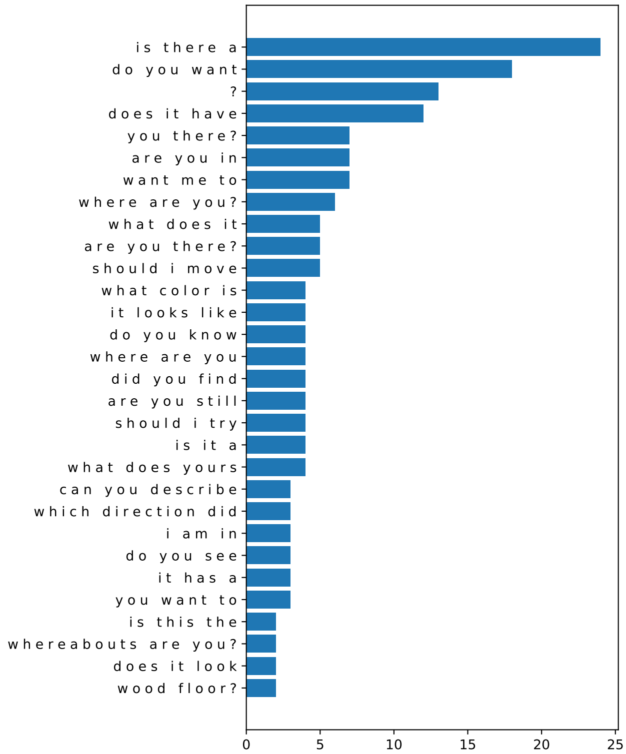

Figure 10: Prefixes of questions (utt.s ending in "?")

the players on the gameboard. Of this, however, only a part is directly accessible for either agent, which is their own current position. The topology of the network must be remembered from experience, if deemed to be relevant. (From observing the actions of the players in the recorded dialogues, it seems unlikely that they attempted to learn the map; they are however able to purposefully return to earlier visited rooms.) More importantly, the current position of the other player is only indirectly observable, through what they report about it. Finally, as we have seen in the examples above, the players often negotiate and agree on a current strategy (e.g., "I find you", "you find me", "we walk around"). As this guides mutual expectations of the players, this is also something that needs to be tracked. On the representation side, we can then assume that an agent will need to track a) their own history of walking through the map (raising interesting questions of how detailed such a representation needs to be or should be made; an artificial agent could help itself by storing the full image for later reference, which would presumably be not enitirely plausible cognitively); b) what has been publicly said and hence could be antecedent to later co-references; c) what they infer about the other player's position; and d) what they assume the current agreed upon strategy is. This clearly is a challenging task; we will in future work first explore hybrid approaches that combine techniques from task-oriented dialogue modelling (Williams and Young, 2007; Buß and Schlangen, 2010) with more recent end-to-end approaches (Côté et al., 2018; Urbanek et al., 2019).

6 Conclusions

We have presented a novel situated dialogue task that brings together visual grounding (talking about objects in a scene), conversational grounding (reaching common ground), and discourse representation (talking about objects that were introduced into the discourse, but aren't currently visible). An agent mastering this task will thus have to combine dialogue processing skills as well as language and vision skills. We hence hope that this task will lead to the further development of techniques that combine both. Our next step is to scale up the collection, to a size where modern machine learning methods can be brought to the task. Besides use in modelling, however, we also think that the corpus can be a valuable resource for linguistic investigations into the phenomenon of negotiating situational grounding.

A Room Types

1. Target room types: bathroom, bedroom, kitchen, basement, nursery, attic, childs_room, playroom, dining_room, home_office, staircase, utility_room, living_room, jacuzzi/indoor, doorway/indoor, locker_room, wine_cellar/bottle_storage, reading_room, waiting_room, balcony/interior
2. Distractor room types: home_theater, storage_room, hotel_room, music_studio, computer_room, street, yard, tearoom, art_studio, kindergarden_classroom, sewing_room, shower, veranda, breakroom, patio, garage/indoor, restroom/indoor, workroom, corridor, game_room, poolroom/home, cloakroom/room, closet, parlor, hallway, reception, carport/indoor, hunting_lodge/indoor
3. Outdoor room types (nodes with a single entry point): garage/outdoor, apartment_building/outdoor, jacuzzi/outdoor, doorway/outdoor, restroom/outdoor, swimming_pool/outdoor, casino/outdoor, kiosk/outdoor, apse/outdoor, carport/outdoor, flea_market/outdoor, chicken_farm/outdoor, washhouse/outdoor, cloister/outdoor, diner/outdoor, kennel/outdoor, hunting_lodge/outdoor, cathedral/outdoor, newsstand/outdoor, parking_garage/outdoor, convenience_store/outdoor, bistro/outdoor, inn/outdoor, library/outdoor

References

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

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR 2018*.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Int. Res.*, 55(1):409–442.

Okko Buß and David Schlangen. 2010. Modelling subutterance phenomena in spoken dialogue systems. In *Proceedings of the 14th International Workshop on the Semantics and Pragmatics of Dialogue (Pozdial 2010)*, pages 33–41, Poznan, Poland.

Donna Byron, Alexander Koller, Jon Oberlander, Laura Stoia, and Kristina Striegnitz. 2007. Generating instructions in virtual environments (give): A challenge and an evaluation testbed for nlg. *Position Papers*, page 3.

Xinlei Chen and C Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2422–2431.

Herbert H. Clark. 1996a. *Using Language*. Cambridge University Press, Cambridge.

Herbert H Clark. 1996b. Using language. 1996. *Cambridge University Press: Cambridge*, pages 274–296.

Herbert H Clark, Susan E Brennan, et al. 1991. Grounding in communication. *Perspectives on socially shared cognition*, 13(1991):127–149.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22:1–39.

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. 2018. TextWorld: A Learning Environment for Text-based Games. *ArXiv*.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2017a. Embodied question answering. *CoRR*, abs/1711.11543.

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

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

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 100–105, Beijing, China. Association for Computational Linguistics.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John Platt, Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In *Proceedings of CVPR*, Boston, MA, USA. IEEE.

Raquel Fernández and David Schlangen. 2007. Referring under restricted interactivity conditions. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 136–139, Antwerp, Belgium.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook Dataset: Building Common Ground through Visually-Grounded Dialogue. In *Proceedings of the 2019 meeting of the Association for Computational Linguistics*, Florence, Italy.

Stevan Harnard. 1990. The symbol grounding problem. *Physica D*, 42:335–346.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 787–798, Doha, Qatar.

Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 36–45, Doha, Qatar. Association for Computational Linguistics.

David Kirsh and Paul Maglio. 1994. On Distinguishing Epistemic from Pragmatic Action. *Cognitive Science*, 18(4):513–549.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163, Denver, Colorado. Association for Computational Linguistics.

Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. 2019. Self-Monitoring Navigation Agent via Auxiliary Progress Estimation. *ArXiv*, pages 1–18.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2015. Generation and comprehension of unambiguous object descriptions. *CoRR*, abs/1511.02283.

Seymour Rosenberg and Bertram D. Cohen. 1964. Speakers' and Listeners' Processes in a Word-Communication Task. *Science*, 145(3637):1201–1204.

Manolis Savva, Angel X. Chang, Alexey Dosovitskiy, Thomas Funkhouser, and Vladlen Koltun. 2017. MINOS: Multimodal indoor simulator for navigation in complex environments. *arXiv:1712.03931*.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. 2019. Habitat: A Platform for Embodied AI Research. *ArXiv*.

David Schlangen. 2016. Grounding, Justification, Adaptation: Towards Machines That Mean What They Say. In *Proceedings of the 20th Workshop on the Semantics and Pragmatics of Dialogue (JerSem)*.

David Schlangen, Tim Diekmann, Nikolai Ilinykh, and Sina Zarrieß. 2018. slurk – A Lightweight Interaction Server For Dialogue Experiments and Data Collection. In *Short Paper Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue (AixDial / semdial 2018)*.

David Schlangen, Sina Zarriess, and Casey Kennington. 2016. Resolving references to objects in photographs using the words-as-classifiers model. In *Proceedings of the 54rd Annual Meeting of the Association for Computational Linguistics (ACL 2016)*.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 721–732, Baltimore, Maryland. Association for Computational Linguistics.

Amanda J Stent. 2011. Computational approaches to the production of referring expressions: Dialog changes (almost) everything. In *PRE-CogSci Workshop*.

Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement Learning*. MIT Press, Cambridge, USA.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *CVPR 2015*, Boston, MA, USA.

Takenobu Tokunaga, Ryu Iida, Asuka Terai, and Naoko Kuriyama. 2012. 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 2012)*.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to Speak and Act in a Fantasy Text Adventure Game. *ArXiv*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition*.

Jason Williams and Steve Young. 2007. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech and Language*, 21(2):231–422.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. *Modeling Context in Referring Expressions*, pages 69–85. Springer International Publishing, Cham.

Sina Zarrieß, Julian Hough, Casey Kennington, Ramesh Manuvinakurike, David DeVault, Raquel Fernandez, and David Schlangen. 2016. Pentoref: A corpus of spoken references in task-oriented dialogues. In *10th edition of the Language Resources and Evaluation Conference*.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.