Learning to Speak and Act in a Fantasy Text Adventure Game

Jack Urbanek[1] **Angela Fan**[1,2] **Siddharth Karamcheti**[1] **Saachi Jain**[1] **Samuel Humeau**[1]
Emily Dinan[1] **Tim Rocktäschel**[1,3] **Douwe Kiela**[1] **Arthur Szlam**[1] **Jason Weston**[1]
[1]Facebook AI Research
[2]LORIA, Nancy
[3]University College London
`light-dms@fb.com`

Abstract

We introduce a large-scale crowdsourced text adventure game as a research platform for studying grounded dialogue. In it, agents can perceive, emote, and act whilst conducting dialogue with other agents. Models and humans can both act as characters within the game. We describe the results of training state-of-the-art generative and retrieval models in this setting. We show that in addition to using past dialogue, these models are able to effectively use the state of the underlying world to condition their predictions. In particular, we show that grounding on the details of the local environment, including location descriptions, and the objects (and their affordances) and characters (and their previous actions) present within it allows better predictions of agent behavior and dialogue. We analyze the ingredients necessary for successful grounding in this setting, and how each of these factors relate to agents that can talk and act successfully.

1 Introduction

There has been remarkable progress in language modeling (Jozefowicz et al., 2016; Devlin et al., 2018; Radford et al., 2019) and building dialogue agents (Dinan et al., 2019a). Nevertheless, the current state of the art uses only the statistical regularities of language data, without explicit understanding of the world that the language describes. This work is built on the hypothesis that dialogue agents embodied in a rich and cohesive (but tractable) world can more easily be trained to use language effectively than those only exposed to standard large-scale text-only corpora.

To that end, we introduce the LIGHT[1] research platform. LIGHT is a multi-player fantasy text adventure world designed for studying situated dialogue, and allows interactions between humans,

models as situated agents, and the world itself. It consists of a large crowdsourced game world (663 locations, 3462 objects and 1755 characters) described entirely in natural language. Within that game world, we collect a large set (11k episodes) of character-driven human-human crowdworker interactions involving actions, emotes, and dialogue, with the aim of training models to engage humans in a similar fashion. Our complete framework is made publicly available in ParlAI (`http://parl.ai/projects/light`).

We use the collected dataset to investigate how a model can both speak *and* act grounded in perception of its environment and dialogue from other speakers. This is done by evaluating state-of-the-art models on our task and evaluating the effects of providing additional grounding. In particular, we adapt the BERT contextual language model (Devlin et al., 2018) to the task of dialogue in two ways: as a bi-ranker, which is fast and practical as a retrieval model, and as a cross-ranker which is slower at inference time but allows more feature cross-correlation between context and response. Both models outperform existing methods. Our ablation analysis shows the importance of each part of the grounding (location, objects, characters, other's actions, self-actions) in terms of the ability to both understand and use language. While models that use grounding show clear improvements, our best performing models are still unable to perform at human level, making our setup a suitable challenge for future research.

2 Related Work

Most recent work in dialogue exploring generative or retrieval models for goal-directed (Henderson et al., 2014; Bordes et al., 2017) or chit-chat tasks (Vinyals and Le, 2015; Sordoni et al., 2015; Zhang et al., 2018) is not situated, or even

---

[1] Learning in Interactive Games with Humans and Text.

grounded in perception. Models typically take the last few utterances from the dialogue history as input, and output a new utterance. While some goal-directed setups may use external knowledge bases (e.g. flight data for airline booking), dialogues tend to implicitly refer to an external world during the conversations without explicit grounding to objects or actions.

Several position papers have proposed virtual embodiment as a strategy for language research (Brooks, 1991; Kiela et al., 2016; Gauthier and Mordatch, 2016; Mikolov et al., 2016; Lake et al., 2017). Single-player text adventure game frameworks for training reinforcement learning agents exist, i.e., Narasimhan et al. (2015) and TextWorld (Côté et al., 2018), but these do not have human dialogue within the game. Similar single player text adventure games have also been used to study referring expressions (Gabsdil et al., 2001, 2002) and parsing (Koller et al., 2004). Yang et al. (2017) and Bordes et al. (2010) also proposed small world setups for instruction following or labeling, but these are much more restricted than the large multi-player text adventure game environment with rich dialogue that we propose here.

A number of visual, rather than text, platforms have been proposed such as House3D (Wu et al., 2018b), HoME (Brodeur et al., 2017), MINOS (Savva et al., 2017), Matterport3D (Chang et al., 2017) and AI2-THOR (Kolve et al., 2017), and the Minecraft MALMO project (Johnson et al., 2016), but they typically are suited to reinforcement learning of actions, and involve templated language for navigation or question answering tasks, if at all (Oh et al., 2017; Yi et al., 2018).

Other examples are instruction-following in the Neverwinter Nights game (Fleischman and Roy, 2005), studies of emotional response in adventure games (Fraser et al., 2018), dialogue about soccer videogames (Pasunuru and Bansal, 2018), placing blocks appropriately given a final plan (Wang et al., 2016) and a more open ended building task using a grid of voxels (Wang et al., 2017). In the latter two cases the communication is one-sided with only the human issuing instructions, rather than dialogue, with the agent only able to act.

There are also setups that consider static language and perception, for example image captioning (Lin et al., 2014), video captioning (Yu et al., 2016), visual QA (Antol et al., 2015) and visual dialogue (Das et al., 2017; Shuster et al., 2018;

Mostafazadeh et al., 2017). While grounded, the agent has no ability to act in these tasks. Talk the Walk (de Vries et al., 2018) introduces a navigation game that involves action, perception and two-way dialogue, but is limited to small grids.

In summary, compared to many setups, our framework allows learning from both actions and (two-way) dialogue, while many existing simulations typically address one or the other but not both. In addition, being based on a gaming setup, our hope is that LIGHT can be fun for humans to interact with, enabling future engagement with our models. All utterances in LIGHT are produced by human annotators, thus inheriting properties of natural language such as ambiguity and coreference, making it a challenging platform for grounded learning of language and actions.

3 LIGHT Environment and Task Setup

LIGHT is a large-scale, configurable text adventure environment for research on learning grounded language and actions. It features both humans and models as agents situated (symbolically) within a multi-player fantasy MUD (multi-user dungeon)-like (Dieterle, 2009) environment. The environment is moderated by a simple game engine which passes dialogue and emote turns between characters and allows actions to cause transitions of the world state. It is in this unimodal (text-only), environment that we consider the agents to be situated.

To facilitate natural human-sourced (fantasy) situations described by natural language, almost the entire environment is crowdsourced, including locations, objects and their affordances, characters and their personalities, and most importantly character interactions: dialogues and actions. These components are collected through a series of annotation tasks that we will now describe. These tasks are designed so that they can be combinatorially recombined. Data quality was maintained by requiring annotators to take a test (see Appendix D). Overall statistics of the collected elements are given in Table 1. This environment can then be used to both train agents, and to evaluate them *in situ* via their online interactions.

Locations We first crowdsourced a set of 663 game location settings from a base set of 37 categories (*countryside*, *forest*, *inside/outside castle*, *shore*, *graveyard*, *bazaar*, . . .– full list in Appendix H) which were selected by us to pro-

Split	Train	Valid	Test Seen	Test Unseen
Locations	589	352	499	74
Objects	2658	1412	1895	844
Characters	1369	546	820	360
Dialogues	8538	500	1000	739
Utterances	110877	6623	13272	9853
Emotes	17609	1156	2495	1301
Actions	20256	1518	3227	1880
Vocabulary Size	32182	11327	11984	9984
Utterance Length	18.3	19.2	19.4	16.2

Table 1: LIGHT dataset statistics.

vide both inspiration and cohesion to annotators. Workers were provided a category and asked to create a description, backstory, names of connected locations, and contained objects and characters. See Table 2a for an example. Many descriptions are quite detailed, and there are clear semantics between entities (e.g. alligators being in swamps, cacti in a desert).

As all remaining tasks build upon the locations created in this first step, we selected 6 location categories (*underwater aquapolis*, *frozen tundra*, *supernatural*, *magical realm*, *city in the clouds*, and *netherworld*) designed to be distinct from the others to provide an isolated set of locations, characters, and objects for testing. These will be used to build what we refer to as an *unseen* test set.

Each location is collected independently, with the eventual aim that they can be glued together as desired to randomize world generation. In this work, we consider actions and dialogues within a single location, so building a world map is not necessary. However, we will show that the environment has considerable influence on the dialogue, actions and grounded learning of models.

Objects We crowdsourced 3462 objects, each with a textual description, and a set of affordances (whether it is a container, can be picked up, has a surface, is a weapon, is wearable, is food, is a drink). See Table 2c for examples. As before, we sourced this list of objects to annotate from the ones annotated for the locations and characters.

Characters We crowdsourced 1755 game characters from animals to trolls and orcs to humans of various types (wizards, knights, village clerk). See Table 2b for detailed examples. Each character has a textual description, a persona (defined as a set of 3-5 profile sentences describing their traits, mod-

eled after the Persona-Chat dataset (Zhang et al., 2018)), and a set of objects that are currently being carried, wielded, or worn. We sourced this list of characters to annotate from the ones provided in the location creation task.

Actions and Emotes There are a set of actions in the game consisting of physical manipulations, and a set of emotes that display feelings to other characters, in line with existing MUDs.

Physical actions include *get*, *drop*, *put*, *give*, *steal*, *wear*, *remove*, *eat*, *drink*, *hug* and *hit*, each taking either one or two arguments, e.g. *put robes in closet*. Every action has an explicit unambiguous effect on the underlying game state, and can only be executed if constraints are met, e.g. if the agent is holding the robes in the latter example. These constraints are what indirectly provide an agent with object affordances, as the list of possible actions provides all ways the agent can interact with their environment (Gibson, 1977).

Emotes include *applaud, blush, cringe, cry, dance, frown . . . , sulk, wave, wink* (22 in total) and have no effect on the game state other than to notify nearby characters of the emote, which can have effects on their behavior. See Appendix E for further detailed descriptions.

Interaction Now that we have a fully realized underlying environment, we can attempt to learn and evaluate agents that can act and speak within it. For this, we collect a human-human dataset of episodic interactions within the environment.

For each dialogue, we place two characters in a random location (either two characters that were already assigned to it, or else randomly assigned characters), complete with the objects assigned to the location and to those characters. Each character has access to their persona, the location description, and the objects present, and the interaction episode begins. The two characters take turns within the episode, and can execute one action (physical action or emote) and produce one dialogue utterance on each turn. We crowdsourced 10,777 dialogues. Examples are given in Figure 1 and Appendix Figures 10-16.

Seen and Unseen Test Sets We provide two distinct test sets. The *seen* test set consists of dialogues set in the same world (set of locations) as the training set, thus also consists of characters, objects, and personas that can appear in the training data. In contrast, the *unseen* test set is com-

Category:	Graveyard
Description:	Two-and-a-half walls of the finest, whitest stone stand here, weathered by the passing of countless seasons. There is no roof, nor sign that there ever was one. All indications are that the work was abruptly abandoned. There is no door, nor markings on the walls. Nor is there any indication that any coffin has lain here... yet.
Backstory:	Bright white stone was all the fad for funerary architecture, once upon a time. It's difficult to understand why someone would abandon such a large and expensive undertaking. If they didn't have the money to finish it, they could have sold the stone, surely - or the mausoleum itself. Maybe they just haven't needed it yet? A bit odd, though, given how old it is. Maybe the gravedigger remembers... if he's sober.
Neighbors:	Dead Tree, south, following a dirt trail behind the mausoleum Fresh Grave, west, walking carefully between fallen headstones
Characters:	gravedigger, *thief, peasant, mouse, bat*
Objects:	wall, *carving, leaf, dirt*

(a) Example room created from the room collection and labelling tasks.

Character:	Thief	Gravedigger
Persona:	I live alone in a tent in the woods. I steal food from the townspeople and coal from the blacksmith. The village police can not find me to put me in jail.	I am low paid labor in this town. I do a job that many people shun because of my contact with death. I am very lonely and wish I had someone to talk to who isn't dead.
Description:	The thief is a sneaky fellow who takes from the people and does so in a way that disturbs the livelihood of the others.	You might want to talk to the gravedigger, specially if your looking for a friend, he might be odd but you will find a friend in him.
Carrying:	meat, potatoes, coal	shovel
Wearing:	dark tunic, cloak	*nothing annotated*
Wielding:	knife	*nothing annotated*

(b) Example characters annotated via character collection tasks.

Object	Description	Tags
shovel	The shovel is made of metal and silver. It is quite sturdy and appears new.	gettable, wieldable
wall	The wall is pure white, the richest of which you have ever seen.	*none*

(c) Example objects annotated via object collection tasks

Table 2: Example entities from the LIGHT environment. Each was collected via tasks described in Section 3.

prised of dialogues collected on the unseen set of locations (described in 3). The unseen test set allows for evaluation of generalization capability to unseen topics in a similar domain and as we shall see, provides a more challenging test for current techniques.

4   Learning Methods

We consider a variety of models that can predict actions, emotes and dialogue, and explore the importance of grounding upon the location, objects, and other characters within the setting. For all models, we represent context as a large text sequence with a special token preceding each input type (persona, setting, self emote, partner emote, etc.). We work with two model classes: *ranking* models that output the maximal scoring response from a set of potential candidate responses and *generative* models that decode word by word.

Ranking Candidates   Each of the three tasks has a different method for determining candidates. Dialogue candidates are the ground truth and 19 randomly chosen candidates. Action candidates are usually the list of all possible actions, however in a no-affordance ablation we provide all well-formed actions over the current environment (which may include things that can't be executed like "wear paint can"). Emote candidates are the 22 possible emotes.

Baseline Ranking Methods   We report a Random baseline (selecting a random candidate from the candidates) and an Information Retrieval (IR) baseline that uses word overlap with TF/IDF weighting. We use *Starspace* (Wu et al., 2018a) which learns a bag-of-words embedding for context and candidates to maximize the inner product of the true label using a ranking loss. Lastly, we use *fastText* (Joulin et al., 2016) to classify which emote should be predicted next as there are only

Persona 1: Servant	Persona 2: King
I come from the lower class. I do what I am told without question. I can not read. I have not seen my family in a long time.	I am a king of the whole empire. I give rules and pursuit them. I am brave and fearless.
Carrying: a duster, a small bucket, a rag Wearing: a shirt	Carrying: a crown, a scepter

Setting: The main foyer is massive. A grand staircase sits to the back of the foyer leading to the upstairs. At the front of the foyer stand two servants ready to help anyone who comes to visit. To the left of the room there is a doorway leading into a corridor. To the right there is a door leading to another corridor for the King's servants. At the foot of the stairs there is a bearskin rug that is staring at you almost as if still hungry. The walls are lined with portraits of the king and his family.

Servant: my humble king. What am I to do to serve you?
King: Ahhh. My loyal servant. Polish my scepter.
  Action: give scepter to servant
Servant: Yes my lord. I will polish it immediately. Am I to return it to you personally?
  Action: put scepter in small bucket
King: Yes. Yes. Of course. Also check the jewels in my crown. They seem loose.
  Action: give crown to servant
Servant: But sire I am not qualified to do that. Would you prefer I take it to someone?
King: Oh fine then.
  Emote: gesture sigh
Servant: I am sorry sir the rug startled me
  Action: drop crown
King: Haha! That's bear I slain on my latest hunting trip. He's a mighty beast!
  Emote: gesture laugh
Servant: and if I may ask where did you go hunting sire?
King: The great woods of course. This bear was stealing children in the kingdom. Surely you heard about it.
Servant: sire. I have not been outside of these walls in quiet some time. I have not seen my family in ages.
King: Such is the life of a servant I suppose. How's that scepter looking?
Servant: it is almost ready sire. and the crown who would you like me to take it to?
  Action: get scepter from small bucket
King: Here just give it back. I'll have the queen find someone.

Figure 1: Example dialogue from the LIGHT dataset.

22 classes. Finally, we compare the performance of our best models to human performance on each of the prediction tasks.

Transformer Memory Network We use the transformer memory-based ranking model from Dinan et al. (2019b). It uses a transformer (Vaswani et al., 2017) to produce separate representations (memory slots) for each sentence from the grounding information (setting, persona, objects). It then performs attention given the dialogue context over the memories to produce a context embedding, which is used to score candidates via the dot product with the transformer-based representation of the candidate. At training time, other samples in the batch are used as negative candidates. For emote prediction, we train by ranking against the full set of possible emotes as there are only 22 distinct classes.

BERT Bi-Ranker and Cross-Ranker We adapt the BERT pretrained language model (Devlin et al., 2018) to the tasks of dialogue and action prediction. We explore two architectures for leveraging BERT. First, we use the *BERT-based Bi-Ranker* to produce a vector representation for the context and a separate representation for each candidate utterance. This representation is obtained by passing the first output of BERT's 12 layers through an additional linear layer, resulting in an embedding of dimension 768. It then scores candidates via the dot product between these embeddings and is trained using a ranking loss.

Second, the *BERT-based Cross-Ranker* instead concatenates the context with each candidate utterance, similar to Wolf et al. (2019). Then, each candidate is scored by computing a softmax over all candidates. Unlike the BERT-based Bi-Ranker, the concatenation of the context with each individual candidate allows the model to attend to the context when encoding each candidate, building a *context-dependent* representation of each candidate. In contrast, the Bi-Ranker can use self-attention to build the candidate and context representations, but cannot modify their representation based upon the context. However, the Cross-Encoder is far more computationally expensive (~11,000 slower than the Bi-Ranker for dialogue retrieval) as each concatenated representation must be recomputed, while the Bi-Ranker can cache the candidates for reuse (see Appendix B).

Query:	chicken	pirate	coffin	rake	tavern	meadow
objects	chicken coop	Pirate swords	the remains	shovel	Ale bottles	flower pot
	eggs	dock	remains	garden	beer	fruit
	a pen for the chickens	cargo	bones	a garden	mug of mead	An enchanted amulet.
	chimney	ship	bones of the innocent	Hand carved stone	a large ornate table	citrus fruit
	corn	seagulls on the dock	adventurer's remains	garden bench	beer keg	fruit trees
characters	chickens	boat captain	spirits of our ancestors	gardener	tavern owner	a deer
	fox trying to steal chickens	captain	mourner	stable hand	bartender	a songbird
	farmers	merchant	zombies	Garden dog	Goblin King's bartender	fruit bats
	The farmers	boat workers	families	stable boy	A serving wench	parent
	farmer	workers	bandit	A stable boy	Serving wench	butterfly
locations	Chicken Pen	Pirate Ship	Old Crypt	Across the King's Garden	The werewolves tavern	Lush meadow
	Corn field	Dock at the Port	sacristy	Hidden garden	Tavern of Browntavia	Flower Field
	Farmer's house	Loading Dock	Disposal area	The garden courtyard	Port Tavern	flower garden
	Large Farm	Fishing Dock	inside temple crypt	Church garden	The bar	Mushroom Hut
	Pig Pen	crew berthing	Sacrifice Chamber	Tool Shed	bazaar outside the royal city	Archery zone
actions	get chicken	hug pirate	put torch in coffin	get rake	hug tavern owner	get flower from meadow
	hug chicken	hit pirate	get torch from coffin	drop Rake	give food item to tavern owner	put flower in Meadow
	hit chicken	steal sword from pirate	put bone in coffin	steal Rake from gardener	give telescope to tavern owner	give Flower to a deer
	give cowbell to chicken	steal cargo from pirate	get bone from coffin	give Rake to thing	drink drink	give Flower to deer
	steal sword from chicken	give cargo to pirate	hit archaeologist	give Rake to person	drop drink	steal Flower from a deer
vocabulary	bock	crew	archaeologist	vegetable	drink	flower
	tasty	ye	robber	carved	drinks	amulet
	bawk	port	crypt	alice	regular	songbird
	moo	sea	loss	hook	item	wasp
	egg	seas	adventures	exorcisms	tip	an

Table 3: Neighboring Starspace phrase embeddings (no pretraining from other data) for different types of entities and actions. The first row are arbitrarily chosen queries (chicken, pirate, coffin, rake, tavern, meadow), and the subsequent rows are their nearest objects, agents, locations, actions and vocabulary in embedding space.

Generative Models Similarly to the ranking setting, we use the Transformer Memory Network from Dinan et al. (2019b) to encode the context features (such as dialogue, persona, and setting). However, to predict an action, emote, or dialogue sequence, we use a Transformer architecture to decode while attending to the encoder output.

For the task of action generation, the set of candidates for ranking models to rank the true action sequence against is constrained by the set of valid actions. For example, the character cannot *pick up book* if there is no book. In the generative model, we compute the log likelihood for the set of possible candidates and normalize to constrain the output space to valid actions to improve the results.

4.1 Implementation

We implement models using PyTorch in ParlAI (Miller et al., 2017). Ranking Transformer models are pretrained on Reddit data (Mazaré et al., 2018) and fine-tuned. We use the BERT (Devlin et al., 2018) implementation provided by Hugging Face[2] with pre-trained weights, then adapted to our Bi-Ranker and Cross-Ranker setups. Generative models are pretrained on the Toronto Books Corpus and fine-tuned except for emote prediction which does not leverage pretraining. We apply byte-pair encoding (Sennrich et al., 2016) to reduce the vocabulary size for generative models. We decode using beam search with beam size 5.

4.2 Evaluation

Automatic To evaluate our models, we calculate percentage accuracy for action and emote prediction. For dialogue, we report Recall@1/20 for ranking the ground truth among 19 other randomly chosen candidates for ranking models and perplexity and unigram F1 for generative models.

Human We present humans with the same ranking task and report R@1/20 to estimate their performance on this task. We report the human accuracy and one standard deviation of error.

Quality control was particularly difficult for human evaluations, as the task of ranking one of 20 candidates is fairly tedious and easy to fake. In order to mitigate these problems, during the evaluation we provide annotated examples on the training in addition to examples on the test set. We only keep the annotations of evaluators who had high accuracy on the training examples to filter low-accuracy evaluators. The training accuracy bar was selected due to the difficulty of the separate tasks as evaluated by our own success rates. Our methods for human evaluation are described in more detail in Appendix F along with how many turns were evaluated.

5 Results

The ranking models are compared in Table 4 on the seen and unseen test sets, and ablations are shown for both the BERT-based Bi-Ranker and

---

[2]https://github.com/huggingface/pytorch-pretrained-BERT

	Test Seen			Test Unseen		
Method	Dialogue R@1/20	Action Acc	Emote Acc	Dialogue R@1/20	Action Acc	Emote Acc
Random baseline	5.0	12.2	4.5	5.0	12.1	4.5
IR baseline	23.7	20.6	7.5	21.8	20.5	8.46
FastText Classification	-	-	13.2	-	-	9.92
Starspace	53.8	17.8	11.6	27.9	16.4	9.8
Transformer MemNet	70.9	24.5	17.3	66.0	21.1	16.6
BERT-based Bi-Ranker	**76.5**	42.5	25.0	70.5	38.6	25.7
BERT-based Cross-Ranker	74.9	**50.7**	**25.8**	69.7	51.8	28.6
Human Performance*	*87.5±2.4	*62.0±3.1	*27.0±2.5	*91.8±1.9	*71.9±3.5	*34.4±2.6

Table 4: Ranking model test performance. (*) Human performance is computed on a subset of data as described in Appendix F.

	Dialogue R@1/20	Action Acc	Emote Acc
BERT-based Bi-Ranker	76.0	38.6	25.1
actions+emotes only	58.6	18.3	10.6
dialogue only	68.1	39.4	23.6
dialogue+action+emote	73.2	40.7	23.1
dialogue+persona	73.3	41.0	26.5
dialogue+setting	70.6	41.2	26.0
dialogue+objects	68.2	37.5	25.5
no objects, no affordances	-	17.6	-

Table 5: BERT-based Bi-Ranker ablations (valid set). The LIGHT environment includes a variety of grounding information: dialogue, action, emote, persona, setting, and object descriptions.

	Dialogue PPL	F1	Action Acc	Emote Acc
Generative Transformer	**27.1**	**13.9**	**13.0**	20.6
actions+emotes only	32.8	9.3	10.5	15.3
dialogue only	28.0	12.5	12.3	20.0
dialogue+action+emote	27.6	12.3	12.8	**22.0**
dialogue+persona	27.8	12.9	12.3	20.8
dialogue+setting	27.8	12.1	11.5	17.8
dialogue+objects	27.7	12.8	11.0	20.2

Table 6: Generative Transformer ablations (valid set).

Generative Transformer in Tables 5 and 6.

5.1 Comparison of Models and Baselines

The IR baseline shows non-random performance, but is outperformed by Starspace which is a stronger baseline. We also tried *FastText* on the emote tasks as classification is appropriate over the 22 possible emotes, and it did better than Starspace. Transformer architectures prove significantly stronger at all tasks, with BERT pre-training proving important for best results as used in the Bi-Ranker and Cross-Ranker architectures. The latter, which can create a context dependent representation of each label candidate, is better at actions and emotes. Human performance is still above all these models, leaving space for future improvements in these tasks. We conducted Wilcoxon signed-rank tests and found comparisons between Starspace and the differing transformers for dialogue on test seen in Table 4 are all significantly different at the p<0.01 level. The generative Transformer model did not work as well as the retrieval models using these metrics.

5.2 Generalization Capability on Unseen Test

The six new unseen test settings are a slightly easier task in absolute numbers (Table 4, right), with improved scores for humans and some models. We observe that BERT-based models exhibit good transfer ability relative to other models, but the gap between their performance and human performance increases from the seen test set to the unseen one. Specifically, there is a 21 point gap on the unseen dialogue test set compared to an 11 point gap on the seen test set, making this a significant challenge for future methods.

5.3 Data Inter-connectedness and Coverage

To illustrate the coverage of entities and actions in the LIGHT world, and the inter-connectedness between them learnable from our data, we trained a simple Starspace embedding model with no pre-built embeddings (so, on our data alone, thus precluding BERT) on all three tasks and show embeddings in Table 3. There is clearly a vast variety of learnable concepts and rich structure between characters, locations, objects, actions and the language describing them. We also show additional t-SNE plots and heatmaps showcasing these relationships in Appendix G.

Persona: I am a part of a group of travelers. I go from town to town selling food to the locals. I grew up poor, but my travels have paid off well.	
Setting 1: Fishmonger's stall, Port A small booth near the edge of the port, it's protected by a piece of old, sun-bleached sailcloth. Baskets of freshly-caught fish, bivalves, and eels sit in the shade in stained wooden troughs of water. A small, aggressive-looking dog is chained to one table, presumably to keep cats away. The stall is redolent with the aroma of fish.	**Setting 2: Dunes, Desert** A massive hilly landscape that is nothing but sand and a few rocks. As you walk this area, you can find some human and animal remains along with broken down wood wagons.
Friend: I wonder what I could eat around here... *Emote:* ponder **Traveler:** Customer, are you here shopping for fish too? **Friend:** What brings you to this place? **Traveler:** I like to come around here for food. Sometimes people who travel through drop the most delicious things. Once in a while it's roasted meet or fish.	**Friend:** I wonder what I could eat around here... *Emote:* ponder **Traveler:** Well, the desert is certainly the wrong place for you my friend. **Friend:** What brings you to this place? **Traveler:** I am travelling to the castle market to sell my goods. I have a terrible sense of direction and have been wondering in the sweltering heat for hours until I found your Oasis.

Table 7: Predicted dialogue by the BERT-based Bi-Ranker (as the *traveler* character) given different settings.

Self name: Sea Witch. Self Previous Dialogue: What do you know about that knight standing over there?		
Input Dialogue + Emote	**Partner**	**Prediction**
His armor is garrish. You know I don't fraternize with land dwellers, *pout*	Mermaid Thief	laugh frown
He is a terrible knight and I hate him, *cry*	Mermaid Troll	scream laugh
I will battle him until the end of my days, *scream*	Mermaid Orc	stare nod

Table 8: Predicted emotes by the Generative Transformer given example inputs from dialogue partner.

5.4 Importance of Grounding

Effect of Various Environment Features We provide a large quantity of information about the environment to each of our models — not only dialogue, but the description of the setting, the character's persona, present objects with descriptions, and more. We analyze the usefulness of the additional grounding information in Tables 5 and 6.

For the dialogue task, having access to all of the environmental information provides the best performance for both retrieval and generative models. Training on dialogue alone substantially decreases performance, while each experiment that adds additional grounding information such as the past actions, persona or the setting description, improves the score. Providing object descriptions as a feature leads to the least improvement. As there are both a large quantity of objects that can be present and objects tend to have long descriptions, it can be challenging for the model to associate such information to a dialogue, action, or emote prediction task. The persona features were found to be impactful, which makes sense as they shape the things the character says (and does).

Action sequence and emote prediction are much improved when using the dialogue history compared to using only past action history. Other features generally have lesser impact in this case, but still give some improvements. Including all features appears challenging for the model, perhaps because of the large input to attend over, resulting in improved results for some ablations.

For the action prediction task, we found that affordance information provides comparable impact to model performance as the rest of the features. A model given only actions and emotes as features and the set of possible actions as candidates has similar perfomance to a model given the best features for the task (dialogue, action, emote, persona, and setting) with a candidate set including actions that would be normally filtered by affordances (such as "wield paint"). This suggests the context or quantity we provided the models in these experiments is not sufficient to be able to implicitly predict affordances at a high enough accuracy to not be distracted by impossible choices.

Most importantly, for all tasks training on the available dialogue data is necessary for good performance. Providing only the action and emote as context results in the worst performance, even on action and emote prediction tasks. Moreover, using dialogue and actions simultaneously improves results almost everywhere. The integrated environment in which agents can both act and speak to other agents provides relevant information that can be used across all tasks.

Context affects predicted utterances We investigate the effect of the environmental context on the predictions by modifying the context and examining the changes in predicted dialogue, ac-

Input from Partner: Wizard	Prediction (Self name: Servant)
I'm feeling sad	hug wizard
You must die!	hit master wizard
Try putting on something else	remove patterned outfit
I'd like you to feed me	give food to master wizard
Can you grab me a paper	give book to wizard's assistant
Can you grab me a beer	get beer
Clean up	get duster
Hide the gold	put gold in satchel
Input from different agents	**Prediction**
Wizard: Can I have some drink?	drop potion
Servant: Can I have some drink?	give wine to servant
Bear: Can I have some drink?	give water to bear

Table 9: Predicted actions by the BERT-based Bi-Ranker given example inputs from the dialogue partner.

tion, and emotes using the BERT-based Bi-Ranker.

The input dialogue and speaker has a strong effect on the predicted action, as shown in Table 9, ranking over all training set actions. For example, a partner asking for an item results in a predicted action dependent on the asker to retrieve it, despite our dataset not being explicitly instructional.

A similar effect is observed for emote prediction. Modifying the dialogue and emote input produces a variety of different predicted emotes in Table 8. Further, keeping the context otherwise fixed but modifying the partner name from *mermaid* to *orc* results in a different predicted emote — the mermaid stating *I will battle him* leads to a *stare* while the orc receives a *nod*.

Finally, for dialogue prediction we find the model produces different outputs that are more appropriate for a given setting, even if the dialogue and characters are the same, see Table 7. With the same text about food, the model retrieved dialogue that was setting appropriate. In the fishmonger's stall, it asked if the human agent was a customer shopping for fish, but in the desert dunes it suggested we might be looking in the wrong place.

6 Conclusion

We introduced a large-scale crowdsourced fantasy text adventure game research platform where agents—both models and humans—can act and speak in a rich and diverse environment of locations, objects, and other characters. We analyzed a variety of models and their ability to leverage the grounding information present in the environment. We hope that this work can enable future research in grounded language learning and further the ability of agents to model a holistic world, complete with other agents within it.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

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

Antoine Bordes, Nicolas Usunier, Ronan Collobert, and Jason Weston. 2010. Towards understanding situated natural language. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 65–72.

Simon Brodeur, Ethan Perez, Ankesh Anand, Florian Golemo, Luca Celotti, Florian Strub, Jean Rouat, Hugo Larochelle, and Aaron Courville. 2017. Home: A household multimodal environment. *arXiv preprint arXiv:1711.11017*.

Rodney A Brooks. 1991. Intelligence without representation. *Artificial intelligence*, 47(1-3):139–159.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*.

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, et al. 2018. Textworld: A learning environment for text-based games. *arXiv preprint arXiv:1806.11532*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Edward Dieterle. 2009. Multi-user virtual environments for teaching and learning. In *Encyclopedia of Multimedia Technology and Networking, Second Edition*, pages 1033–1041. IGI Global.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019a. The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019b. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Michael Fleischman and Deb Roy. 2005. Intentional context in situated natural language learning. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 104–111. Association for Computational Linguistics.

Jamie Fraser, Ioannis Papaioannou, and Oliver Lemon. 2018. Spoken conversational ai in video games–emotional dialogue management increases user engagement. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 179–184. ACM.

Malte Gabsdil, Alexander Koller, and Kristina Striegnitz. 2001. Building a text adventure on description logic. In *International Workshop on Applications of Description Logics, Vienna, September*, volume 18.

Malte Gabsdil, Alexander Koller, and Kristina Striegnitz. 2002. Natural language and inference in a computer game. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Jon Gauthier and Igor Mordatch. 2016. A paradigm for situated and goal-driven language learning. *arXiv preprint arXiv:1610.03585*.

James J Gibson. 1977. The theory of affordances. *Hilldale, USA*, 1(2).

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. 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.

Matthew Johnson, Katja Hofmann, Tim Hutton, and David Bignell. 2016. The malmo platform for artificial intelligence experimentation. In *IJCAI*, pages 4246–4247.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Douwe Kiela, Luana Bulat, Anita L Vero, and Stephen Clark. 2016. Virtual embodiment: A scalable long-term strategy for artificial intelligence research. *arXiv preprint arXiv:1610.07432*.

Alexander Koller, Ralph Debusmann, Malte Gabsdil, and Kristina Striegnitz. 2004. Put my galakmid coin into the dispenser and kick it: Computational linguistics and theorem proving in a computer game. *Journal of Logic, Language and Information*, 13(2):187–206.

Eric Kolve, Roozbeh Mottaghi, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. 2017. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing data using t-sne.

P.-E. Mazaré, S. Humeau, M. Raison, and A. Bordes. 2018. Training Millions of Personalized Dialogue Agents. *ArXiv e-prints*.

Tomas Mikolov, Armand Joulin, and Marco Baroni. 2016. A roadmap towards machine intelligence. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 29–61. Springer.

A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. 2017. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv preprint arXiv:1701.08251*.

Karthik Narasimhan, Tejas Kulkarni, and Regina Barzilay. 2015. Language understanding for text-based games using deep reinforcement learning. *arXiv preprint arXiv:1506.08941*.

Junhyuk Oh, Satinder Singh, Honglak Lee, and Pushmeet Kohli. 2017. Zero-shot task generalization with multi-task deep reinforcement learning. *arXiv preprint arXiv:1706.05064*.

Ramakanth Pasunuru and Mohit Bansal. 2018. Game-based video-context dialogue. *arXiv preprint arXiv:1809.04560*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL*.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2018. Engaging image chat: Modeling personality in grounded dialogue. *arXiv preprint arXiv:1811.00945*.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *Proceedings of the 31st International Conference on Machine Learning, Deep Learning Workshop*, Lille, France.

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

Sida I Wang, Samuel Ginn, Percy Liang, and Christoper D Manning. 2017. Naturalizing a programming language via interactive learning. *arXiv preprint arXiv:1704.06956*.

Sida I Wang, Percy Liang, and Christopher D Manning. 2016. Learning language games through interaction. *arXiv preprint arXiv:1606.02447*.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Ledell Yu Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes, and Jason Weston. 2018a. Starspace: Embed all the things! In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. 2018b. Building generalizable agents with a realistic and rich 3d environment. *arXiv preprint arXiv:1801.02209*.

Zhilin Yang, Saizheng Zhang, Jack Urbanek, Will Feng, Alexander H Miller, Arthur Szlam, Douwe Kiela, and Jason Weston. 2017. Mastering the dungeon: Grounded language learning by mechanical turker descent. *arXiv preprint arXiv:1711.07950*.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems*, pages 1039–1050.

Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. 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*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.