Nerea Losada and Jesús Calleja

Learning to Speak and Act in a Fantasy Text Adventure Game

Abstract

We explain the process of creation of an environment, LIGHT, where agents can be tested. The environment is based in a text adventure game where agents can perceive, emote and act and in the meantime talk to other agents. The features of the environment have been crowdsourced. The most important part about the research is about grounded learning. Agents are able to use the underlying information of the environment to correctly predict dialogs, emotions, actions, etc.

Introduction

The models that are currently used don't exploit the totality of the language, limiting to only the statistical data and ignoring what the language itself describes. To understand the world that the language describes, they introduce a multi-player fantasy text adventure world designed for studying dialogue, and allows interactions between humans, agents, and the world itself. The environment consists of a large crowdsourced game world (663 locations, 3462 objects and 1755 characters) described entirely in natural language.

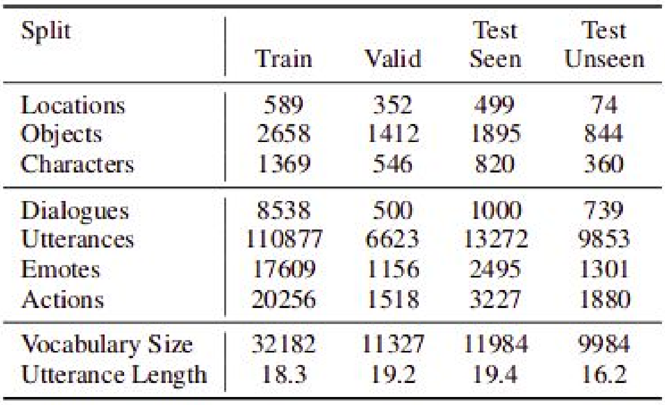
Throughout all the data registered, they have collected 11.000 episodes worth of human-human interactions with all the elements needed to create the environment, such as dialogs and actions. This data is useful to feed it to the agents so that they can learn properly. They use the information of the location, character they are interacting with as well as past dialogue between them that can serve as context. In the report, they have used the BERT model and adapted it to understand the information of the text game. The results have been successful as they have outperformed the previous best results. They have also completed an analysis that shows how important each part of the environment is so that the agent can understand and use the language. Models still haven't surpassed the human level, so there is room for improvement.

Related work

In summary, compared to many setups, LIGHT framework allows learning from both actions and two-way dialogue, while many existing simulations typically address one or the other but not both. The language LIGHT has been created on is based on the crowdworkers that participated, thus inheriting properties of natural language such as ambiguity and coreference, making it a challenging platform for grounded learning of language and actions.

Description of the dataset

In the next table, we can observe the statistics of the collected elements in the LIGHT environment. As we can see, there are locations, objects, characters, dialogues, utterances, emotes and actions. We can also appreciate the vocabulary size and the utterance length. The data is split into four sets: train, valid, test seen and test unseen.



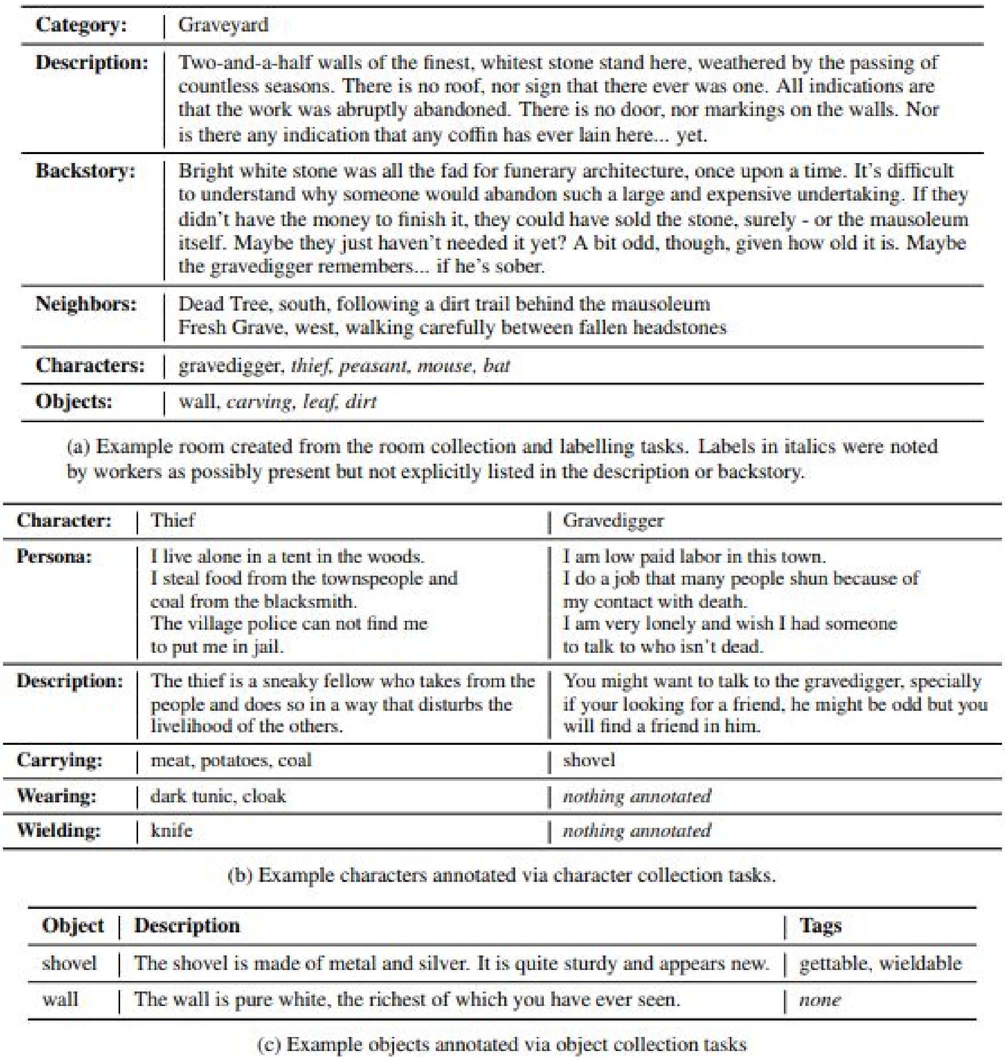
Split	Train	Valid	Test Seen	Test Unseen
Locations	589	352	499	74
Objects	2658	1412	1895	844
Characters	1369	546	820	360
Dialogues	8538	500	1000	739
Utterances	110877	6623	13272	9853
Emotes	17609	1156	2495	1301
Actions	20256	1518	3227	1880
Vocabulary Size	32182	11327	11984	9984
Utterance Length	18.3	19.2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

Description of LIGHT environment

The environment used is LIGHT, which is a large-scale and configurable text adventure environment. By using it, they can research on learning grounded language and actions. It features both humans and models as integrated agents by a multi-player fantasy MUD (multi-user dungeon), which is a multiplayer real-time virtual world, usually text-based. They combine elements of role-playing games, hack and slash, player versus player, interactive fiction and online chat.

The whole environment is crowdsourced, including locations, objects and their affordances, characters and their personalities, dialogues and actions. This is because this way is easier to get natural human-sourced situations described by natural language. To obtain those components, they use a series of annotation tasks described in the next items.

This environment can then be used to both train agents and to evaluate them in situ via their online interactions.



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 ever 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. Labels in italics were noted by workers as possibly present but not explicitly listed in the description or backstory.

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

Location

From a base set in which they are 37 categories chosen specifically to supply inspiration and cohesion to annotators, such as countryside, forest, inside/outside castle, shore, graveyard, bazaar… They crowdsourced a set of 663 game location settings.

They gave workers a category and requested to create a description, backstory, names of connected locations and contained objects and characters. Since the descriptions are quite precise, there are clear semantics between individuals.

Then, they selected 6 location categories: underwater aquapolis, frozen tundra, supernatural, magical realm, city in the clouds and netherworld, which were created with the aim of making a distinction between those and the others to offer an isolated set of locations, characters, and objects for testing. This way, those will be used to build an unseen test set, which is a set that contains dialogues collected on the unseen set of locations. This set permits to evaluate the generalization capability to unseen topics in a similar domain and it also provides a more difficult test for state-of-the-art techniques.

In addition, each location is collected alone, with the objective of pasting them together as wanted to randomize world generation.

In this work, they contemplate actions and dialogues within a single location, this way is not necessary to build a world map. Despite this, they show that the dialogue, actions and grounded learning of models depend on the environment, which has much influence on them.

Characters

They crowdsourced 1755 game characters, which are animals, trolls and orcs and humans of different types, such as wizards, knights and village clerk. Each of them is described by a textual description, has a persona defined as a set of 3-5 profile sentences and a set of objects that they carry, use or wear. They sourced this list of characters to annotate from the ones annotated for the locations.

Objects

They crowdsourced 3462 objects, each with a textual description, as characters, and a set of actions that can be performed with them, whether it is a container, can be picked up, has a surface, is a weapon, is wearable, is food, is a drink, etc. As before, they sourced this list of objects to annotate from the ones annotated for the locations and characters.

Actions and Emotes

There are a set of actions and a set of emotes in the game. The first ones consist of physical manipulations, and the second ones show feelings to other characters.

The physical actions that are included in the set are 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.

Emotes include applaud, blush, cringe, cry, dance, frown, sulk, wave, wink... (22 in total). Those have no effect on the game state, nevertheless they are used to notify nearby characters of the emote, which can have effects on their behavior.

Interaction

They try to learn and evaluate agents that can act and speak within it. To do this, they collect data from interactions between two humans within the environment. To get information about those interactions, they place two characters in a random location complete with the objects attributed to the location and to the characters. This way, the interaction episode begins and the two characters take turns to execute either one physical action or an emote and produce one dialogue.

Seen and Unseen Test Sets

There are two different test sets: the seen test set and the unseen test set. On the one hand, the seen test set includes dialogues set in the same set of locations, which are used as training data. Besides, it consists of characters, objects and personas that can be in the training set. On the other hand, the unseen test set has dialogues collected on the unseen set of locations. This way, this test set serves to evaluate the generalization capability to unseen topics in similar domain, due to it provides a more difficult test for current techniques, as explained before.

Learning Methods

The goal of the experiment is to show the importance of grounding learning. For that purpose, the training data has been labeled with the type of information that gives to the context (self emotion, partner's emotion, setting…). For training models, they considered two options: a ranking model, which outputs the result with the best score, and a generative one.

The ranking models that have been used are the following ones: Transformer Memory Network , BERT Bi-Ranker and BERT Cross Ranker. The first one is the generative model and the latter ones are ranking models.[1]

As well as the model, for comparison purposes, some baselines have been introduced. First, we have a random baseline, which is basically is selecting a candidate by chance, an information retrieval baseline, Starspace, and FastText for emotions.

The implementation of the models used the PyTorch library as well as the ParlAI framework. The models were pretrained with some corpuses from the internet, such as Reddit data, and fine-tuned.

Evaluation

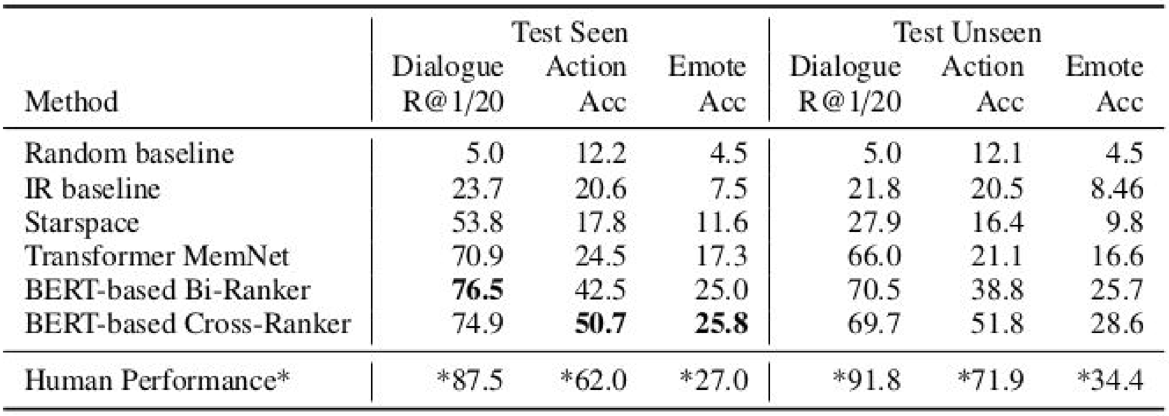
Automatic	Human
To evaluate their models, theye calculate percentage accuracy for action and emote prediction. For dialogue, they randomly choose other 19 candidates and report Recall@1/20 for ranking the ground truth among them. This way they ranked models and perplexity and unigram F1 for generative models.	They present humans with the same ranking task and report R@1/20 to estimate their performance on this task. For the evaluation, they provide annotated examples on the training besides to examples on the test set. They only keep those who had high accuracy on the training examples to filter low accuracy evaluators. Depending on the difficulty of the separate

[1] We won't go deep in the explanation of the models, as this report emphasize other topics.

	tasks they selected the training accuracy bar.

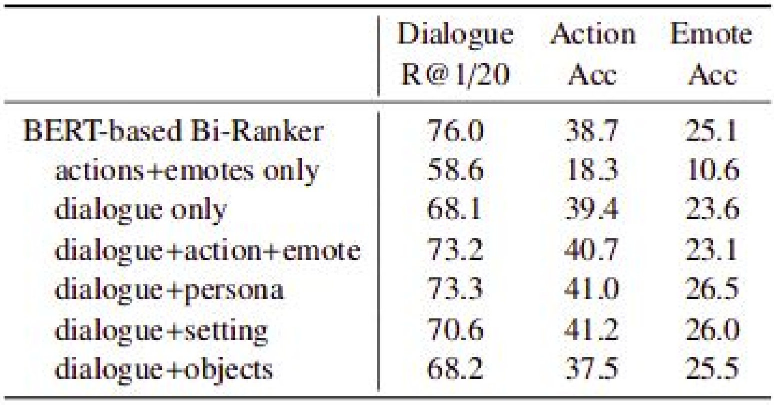
Results

Compared to the baselines, the Transformer based models show better results, but they are far from being on par with human performance.



	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
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	38.8	25.7
BERT-based Cross-Ranker	74.9	**50.7**	**25.8**	69.7	51.8	28.6
Human Performance*	*87.5	*62.0	*27.0	*91.8	*71.9	*34.4

As we said before, grounding information was key to improve the performance of the models. When grounding information is given to the model, this one shows the best performance. It gets worse when there is no additional apart from the dialogue.
We can also see that persona information is the most important one in the dialogue task. Anyway, we get the best result combining all the available information. We only show the Bi-Ranker model because it’s the one that got the best accuracy score.



	Dialogue R@1/20	Action Acc	Emote Acc
BERT-based Bi-Ranker	76.0	38.7	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

Our experience

For us, it has been an interesting task to search and read about the learning to speak and act in a fantasy text adventure game. We have learnt how text games work and which is the process they follow to reach their objective, this is, to get the characters learn how to speak and act.

We find this project fascinating, due to we had the opportunity of trying the demo and seeing how it really works, and which is the result they have obtained.

Conclusion

To finish and conclude this report, we could sum up saying that they analyzed a variety of models and their ability to take advantage of the grounding information right in the environment. For that aim, they introduced LIGHT, the mentioned large-scale crowdsourced fantasy text adventure game research platform. By that platform, agents (both models and humans) can act and speak in a rich and diverse environment of locations, objects, and other characters.

They hope that this work will be useful in the future to enable research in grounded language learning and the ability of agents to deal with a whole world, complete with other agents within it.

PAPER: https://arxiv.org/abs/1903.03094
PROJECT URL: http://parl.ai/projects/light/