A Survey of Reinforcement Learning Informed by Natural Language

Jelena Luketina[1*] , **Nantas Nardelli**[1,2] , **Gregory Farquhar**[1,2] , **Jakob Foerster**[2] ,
Jacob Andreas[3] , **Edward Grefenstette**[2,4] , **Shimon Whiteson**[1] and **Tim Rocktäschel**[2,4*]

[1]University of Oxford
[2]Facebook AI Research
[3]Massachusetts Institute of Technology
[4]University College London

Abstract

To be successful in real-world tasks, Reinforcement Learning (RL) needs to exploit the compositional, relational, and hierarchical structure of the world, and learn to transfer it to the task at hand. Recent advances in representation learning for language make it possible to build models that acquire world knowledge from text corpora and integrate this knowledge into downstream decision making problems. We thus argue that the time is right to investigate a tight integration of natural language understanding into RL in particular. We survey the state of the field, including work on instruction following, text games, and learning from textual domain knowledge. Finally, we call for the development of new environments as well as further investigation into the potential uses of recent Natural Language Processing (NLP) techniques for such tasks.

1 Introduction

Languages, whether natural or formal, allow us to encode abstractions, to generalize, to communicate plans, intentions, and requirements, both to other parties and to ourselves [Gopnik and Meltzoff, 1987]. These are fundamentally desirable capabilities of artificial agents. However, agents trained with traditional approaches within dominant paradigms such as Reinforcement Learning (RL) and Imitation Learning (IL) typically lack such capabilities, and struggle to efficiently learn from interactions with rich and diverse environments. In this paper, we argue that the time has come for natural language to become a first-class citizen of solutions to sequential decision making problems (*i.e.* those often approached with RL[1]). We survey recent work and tools that are beginning to make this shift possible, and outline next research steps.

Humans are able to learn quickly in new environments due to a rich set of commonsense priors about the world [Spelke and Kinzler, 2007], some of which are reflected in natural language [Shusterman *et al.*, 2011; Tsividis *et al.*, 2017]. It is

thus reasonable to question whether agents can learn not only from rewards or demonstrations, but also from information encoded using language, in order to improve generalization and sample efficiency—especially when (a) learning is severely constrained by data efficiency due to limited or expensive environment interactions, and where (b) human priors that would help to solve the task are or can be expressed easily in natural language. Furthermore, many, if not most, real-world tasks require agents to process language by design, whether to enable interaction with humans, or when using existing interfaces and knowledge bases. This suggests that the use of language in RL has both broad and important applications.

Information contained in both generic and task-specific large textual corpora may be highly valuable for decision making. Pre-training neural representations of words and sentences from large generic corpora has already shown great success in transferring syntactic and, to some extent, semantic information to downstream tasks in natural language understanding [Peters *et al.*, 2018b; Goldberg, 2019; Tenney *et al.*, 2019]. Cross-domain transfer from self-supervision on generic language data might similarly help to initialize RL agents. Task-specific corpora like wikis or game manuals could be leveraged by machine reading techniques [Banko *et al.*, 2007] to inform agents of valuable features and policies [Eisenstein *et al.*, 2009; Branavan *et al.*, 2012], or task-specific environmental dynamics and reward structures [Narasimhan *et al.*, 2018; Bahdanau *et al.*, 2019].

Previous attempts at using language for RL tasks in these ways have mostly been limited to relatively small corpora [Janner *et al.*, 2018], or synthetic language [Hermann *et al.*, 2017]. We argue that recent advances in representation learning [Peters *et al.*, 2018a; Devlin *et al.*, 2018; Radford *et al.*, 2019] make it worth revisiting this research agenda with a much more ambitious scope. While the problem of grounding (*i.e.* learning the correspondence between language and environment features) remains a significant research challenge, past work has already shown that high-quality linguistic representations can assist cross-modal transfer outside the context of RL (*e.g.* using semantic relationships between labels to enable zero-shot transfer in image classification [Frome *et al.*, 2013; Socher *et al.*, 2013]).

We first provide background on RL and techniques for self-supervision and transfer in NLP (§2). We then review prior work, considering settings where interaction with language is
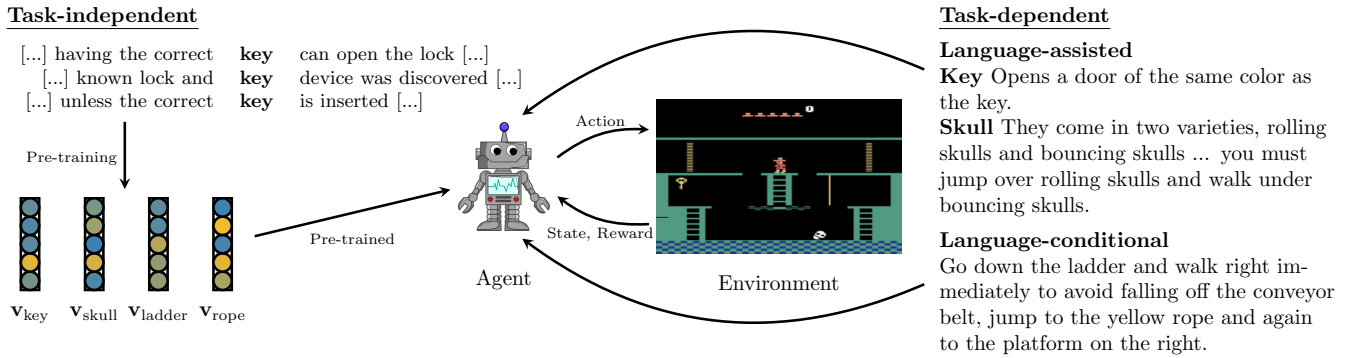
---

[*]Contact authors: jelena.luketina@cs.ox.ac.uk, rockt@fb.com.

[1]We write RL for brevity and to focus on a general case, but our arguments are relevant for many sequential decision making approaches, including IL and planning.

Task-independent

[...] having the correct **key** can open the lock [...]
[...] known lock and **key** device was discovered [...]
[...] unless the correct **key** is inserted [...]

Pre-training

$\mathbf{v}_{key}$ $\mathbf{v}_{skull}$ $\mathbf{v}_{ladder}$ $\mathbf{v}_{rope}$

Action

Pre-trained

State, Reward

Agent   Environment

Task-dependent

Language-assisted
Key Opens a door of the same color as the key.
Skull They come in two varieties, rolling skulls and bouncing skulls ... you must jump over rolling skulls and walk under bouncing skulls.

Language-conditional
Go down the ladder and walk right immediately to avoid falling off the conveyor belt, jump to the yellow rope and again to the platform on the right.

Figure 1: Illustration of different roles and types of natural language information in reinforcement learning. We differentiate between *language-conditional* setting in which language is a part of the task formulation (*e.g.* natural language instructions that specify the goal or reward), and *language-assisted* setting where language information is not necessary to solve the task but can assists learning (*e.g.* by providing information about the environment dynamics). The language information itself can be *task-dependent*, *i.e.* specific to the task such as tutorials or instructions, or *task-independent*, for instance, conveying general priors about the world through pre-trained language representations.

necessary (§3.1) and where language can optionally be used to facilitate learning (§3.2). In the former category we review instruction following, induction of reward from language, and environments with text in the action or observation space, all of which have language in the problem formulation itself. In the latter, we review work that uses language to facilitate RL by transfer from domain-specific textual resources, or as a means of representing policies.

We conclude by identifying what we believe are the most important challenges for integrating natural language in RL (§4). Inspired by gaps in the existing literature, we advocate the development of new research environments utilizing domain knowledge in natural language, as well as a wider use of NLP methods such as pre-trained language models and parsers to inform RL agents about the structure of the world.

2 Background

2.1 Reinforcement and Imitation Learning

Reinforcement Learning [Sutton and Barto, 2018] is a framework that enables agents to reason about sequential decision making as an optimization process. Problems are formulated as Markov Decision Processes (MDPs), tuples $\langle S, A, T, R, \gamma \rangle$ where S is the set of states, A the set of actions, T the transition probability function $T : S \times A \times S \rightarrow [0, 1]$, R the reward function $R : S \times A \times S \rightarrow \mathbb{R}$, and $\gamma \in [0, 1)$ is a discount factor, typically set by either the environment or the agent designer. Given this setup, the goal of the optimization process is to find a policy $\pi(a|s) = p(A = a|S = s)$ that maximizes the expected discounted cumulative return $\sum_{k=0}^{\infty} \gamma^k r_{k+1}$. This framework is also used in Imitation Learning (IL), a setting in which the rewards are not observed, but the learning algorithm has access to a set of trajectories under optimal or sub-optimal policies. IL methods can then find approximations of the optimal policy, which can be used as a form of initialization, auxiliary objective, or for value estimation.

Since their inception, RL algorithms have been successful in applications such as continuous control [White and Sofge, 1992], dialogue systems [Singh *et al.*, 2002], and board

games [Tesauro, 1995]. Recent improvements in function approximation and pattern recognition made possible by deep learning have allowed RL to scale to problems with high dimensional input spaces such as video games [Torrado *et al.*, 2018] and complex planning problems such as Go [Silver *et al.*, 2017]. Nonetheless, these methods remain sample inefficient, requiring millions or billions of interactions, and often generalize poorly to tasks only slightly different from those seen during training. This severely limits the use of RL for real-world tasks. See [Sutton and Barto, 2018] for a comprehensive introduction to RL and to [Arulkumaran *et al.*, 2017] and [Osa *et al.*, 2018] for reviews on recent algorithmic developments.

2.2 Transfer from Natural Language

NLP has seen a recent surge of models that transfer syntactic and semantic knowledge to various downstream tasks. Current NLP systems commonly employ deep learning models and embed (sequences of) words using dense vector representations. These vector representations are often pre-trained from large textual corpora and fine-tuned for a given task. Common techniques learn individual word representations from co-occurrence statistics [Deerwester *et al.*, 1990; Mikolov *et al.*, 2013] or contextual word-representations using (pseudo) language model objectives [Peters *et al.*, 2018a; Devlin *et al.*, 2018]. Both classes of models are motivated by Firth's distributional hypothesis ("You shall know a word by the company it keeps") [Firth, 1957], which suggests that the learned vector representation of a word like 'scorpion' should be similar to 'spider' if the corresponding words appear in similar contexts, *e.g.*, if they can both be found around other words like 'venomous' or 'exoskeleton'.

Such (contextual) word representations can transfer knowledge to downstream tasks that have to deal with language as, for example, in [Socher *et al.*, 2013; Frome *et al.*, 2013; Howard and Ruder, 2018; Peters *et al.*, 2018b; Goldberg, 2019; Tenney *et al.*, 2019], to name just a few. For instance, consider a text classification problem where we are tasked with assigning a document containing the word 'scorpion' to the

topic 'arachnids' even though we have only observed 'spider' during training.

World and task-specific knowledge communicated in natural language could be similarly transferred to sequential decision making problems. For instance, learning agents can benefit from understanding explicit goals ("go to the door on the far side of the room"), constraints on policies ("avoid the scorpion"), or generic information about the reward or transition function ("scorpions are fast"). Furthermore, pre-trained language models could play an important role in transferring world knowledge such as object affordances ("a key is used for [opening doors | unlocking boxes | investigating locked chests]"). Similarly to recent question-answering [Chen *et al.*, 2017a] and dialog systems [Dinan *et al.*, 2018], agents could learn to make use of information retrieval and NLP components to actively seek information required for making progress on a given task [Branavan *et al.*, 2012].

3 Current Use of Natural Language in RL

In reviewing efforts that integrate language in RL we highlight work that develops tools, approaches, or insights that we believe may be particularly valuable for improving the generalization or sample efficiency of learning agents through the use of natural language. As illustrated in Figure 1, we separate the literature into **language-conditional** RL (in which interaction with language is necessitated by the problem formulation itself) and **language-assisted** RL (in which language is used to facilitate learning). The two categories are not mutually exclusive, in that for some language-conditional RL tasks, NLP methods or additional textual corpora are used to assist learning [Bahdanau *et al.*, 2019; Goyal *et al.*, 2019].

To easily acquire data and constrain the difficulty of problems considered, the majority of these works use synthetic language (automatically generated from a simple grammar and limited vocabulary) rather than language generated by humans. These often take the form of simple templates, *e.g.* "what colour is ¡object¿ in ¡room¿", but can be extended to more complex templates with relations and multiple clauses [Chevalier-Boisvert *et al.*, 2019].

3.1 Language-conditional RL

We first review literature for tasks in which integrating natural language is unavoidable, *i.e.*, when the task itself is to interpret and execute instructions given in natural language, or natural language is part of the state and action space. We argue in (§4.1) that approaches to such tasks can also be improved by developing methods that enable transfer from general and task-specific textual corpora. Methods developed for language-conditional tasks are relevant for language-assisted RL as they both deal with the problem of grounding natural language sentences in the context of RL. Moreover, in tasks such as following sequences of instructions, the full instructions are often not necessary to solve the underlying RL problem but they assist learning by structuring the policy [Andreas *et al.*, 2017] or by providing auxiliary rewards [Goyal *et al.*, 2019].

Instruction Following

Instruction following agents are presented with tasks defined by high-level (sequences of) instructions. We focus on instructions that are represented by (at least somewhat natural) language, and may take the form of formal specifications of appropriate actions, of goal states (or goals in general), or of desired policies. Effective instruction following agents execute the low level actions corresponding to the optimal policy or reach the goal specified by their instructions, and can generalize to unseen instructions during testing.

In a typical instruction following problem, the agent is given a description of the goal state or of a preferred policy as a proxy for a description of the task [MacMahon *et al.*, 2006; Kollar *et al.*, 2010]. Some work in this area focuses on simple object manipulation tasks [Wang *et al.*, 2016; Bahdanau *et al.*, 2019], while other work focuses on 2D or 3D navigation tasks where the goal is to reach a specific entity. Entities might be described by predicates ("Go to the red hat") [Hermann *et al.*, 2017; Chaplot *et al.*, 2018] or in relation to other entities ("Reach the cell above the western-most rock.") [Janner *et al.*, 2018; Chen *et al.*, 2018]. Earlier approaches use object-level representation and relational modeling to exploit the structure of the instruction in relation to world entities, parsing the language instruction into a formal language [Kuhlmann *et al.*, 2004; Chen and Mooney, 2011; Artzi and Zettlemoyer, 2013; Andreas and Klein, 2015]. More recently, with the developments in deep learning, a common approach has been to embed both the instruction and observation to condition the policy directly [Mei *et al.*, 2016; Hermann *et al.*, 2017; Chaplot *et al.*, 2018; Janner *et al.*, 2018; Misra *et al.*, 2017; Chen *et al.*, 2018]. Human-generated natural language instructions are used in [MacMahon *et al.*, 2006; Bisk *et al.*, 2016; Misra *et al.*, 2017; Janner *et al.*, 2018; Chen *et al.*, 2018; Anderson *et al.*, 2018; Goyal *et al.*, 2019; Wang *et al.*, 2018]. Due to data-efficiency limitations of RL, this is not a standard in RL-based research [Hermann *et al.*, 2017].

The line of work involving sequences of instructions has strong ties to Hierarchical RL [Barto and Mahadevan, 2003], with individual sentences or clauses from instructions corresponding to subtasks [Branavan *et al.*, 2010]. When the vocabulary of instructions is sufficiently simple, an explicit options policy can be constructed that associates each task description with its own modular sub-policy [Andreas *et al.*, 2017]. A more flexible approach is to use a single policy that conditions on the currently executed instruction, allowing some generalization to unseen instructions [Mei *et al.*, 2016; Oh *et al.*, 2017]. However, current approaches of this form require first pre-training the policy to interpret each of the primitives in a single-sentence instruction following setting. The internal compositional structure of instructions can then be exploited in various ways. For example, [Oh *et al.*, 2017] achieve generalization to unseen instructions by forcing instruction embeddings to capture analogies, *e.g.*, [Visit,X] : [Visit,Y] :: [Pick up,X] : [Pick up,Y].

Rewards from Instructions

Another use of instructions is to induce a reward function for RL agents or planners to optimize. This is relevant when

the environment reward is not available to the agent at test time, but is either given during training [Tellex *et al.*, 2011] or can be inferred from (parts of) expert trajectories. In order to apply instruction following more broadly, there needs to be a way to automatically evaluate whether the task specified by the instruction has been completed. The work addressing this setting is influenced by methods from the inverse reinforcement learning (IRL) literature [Ziebart *et al.*, 2008; Ho and Ermon, 2016]. A common architecture consists of a reward-learning module that learns to ground an instruction to a (sub-)goal state or trajectory segment, and is used to generate a reward for a policy-learning module or planner.

When full demonstrations are available, the reward function can be learned using standard IRL methods like MaxEnt IRL [Ziebart *et al.*, 2008] as in [Fu *et al.*, 2019], or maximum likelihood IRL [Babes *et al.*, 2011] as in [MacGlashan *et al.*, 2015], who also learn a joint generative model of rewards, behaviour, and language. Otherwise, given a dataset of goal-instruction pairs, as in [Bahdanau *et al.*, 2019], the reward function is learned through an adversarial process similar to that of [Ho and Ermon, 2016]. For a given instruction, the reward-learning module aims to discriminate goal states from the states visited by the policy (assumed non-goal), while the agent is rewarded for visiting states the discriminator cannot distinguish from the goal states.

When environment rewards are available but sparse, instructions may still be used to generate auxiliary rewards to help learn efficiently. In this setting, [Goyal *et al.*, 2019] and [Wang *et al.*, 2018] use auxiliary reward-learning modules trained offline to predict whether trajectory segments correspond to natural language annotations of expert trajectories. [Agarwal *et al.*, 2019] perform a meta-optimisation to learn auxiliary rewards conditioned on features extracted from instructions. The auxiliary rewards are learned so as to increase performance on the true objective after being used for a policy update. As some environment rewards are available, these settings are closer to language-assisted RL.

Language in the Observation and Action Space

Environments that use natural language as a first-class citizen for driving the interaction with the agent present a strong challenge for RL algorithms. Using natural language requires common sense, world knowledge, and context to resolve ambiguity and cheaply encode information [Mey, 1993]. Furthermore, linguistic observation and action spaces grow combinatorially as the size of the vocabulary and the complexity of the grammar increase. For instance, compare the space of possible instructions when following cardinal directions (*e.g.* "go north") with reaching a position that is described in relative terms (*e.g.* "go to the blue ball south west of the green box").

Text games, such as Zork [Infocom, 1980], are easily framed as RL environments and make a good testbed for structure learning, knowledge extraction, and transfer across tasks [Branavan *et al.*, 2012]. [DePristo and Zubek, 2001; Narasimhan *et al.*, 2015; Yuan *et al.*, 2018] observe that when the action space of the text game is constrained to verb-object pairs, decomposing the Q-function into separate parts for verb and object provides enough structure to make learning more tractable. However, they do not show how to scale this ap-

proach to action-sentences of arbitrary length. To facilitate the development of a consistent set of benchmarks in this problem space, [Côté *et al.*, 2018] propose *TextWorld*, a framework that allows the generation of instances of text games that behave as RL environments. They note that existing work on word-level embedding models for text games (e.g. [Kostka *et al.*, 2017; Fulda *et al.*, 2017]) achieve good performance only on easy tasks.

Other examples of settings where agents are required to interact using language include *dialogue systems* and *question answering* (Q&A). The two have been a historical focus in NLP research and are extensively reviewed by [Chen *et al.*, 2017b] and [Bouziane *et al.*, 2015] respectively. Recent work on visual Q&A (VQA) have produced a first exploration of multi-modal settings in which agents are tasked with performing both visual and language-based reasoning [Antol *et al.*, 2015; Johnson *et al.*, 2017; Massiceti *et al.*, 2018]. Embodied Q&A (EQA) extends this setting, by requiring agent to explore and navigate the environment in order to answer queries [Das *et al.*, 2018a; Gordon *et al.*, 2018], for example, "How many mugs are in the kitchen?" or "Is there a tomato in the fridge?". By employing rich 3D environments, EQA tasks require agents to carry out multi-step planning and reasoning under partial observability. However, since in the existing work the agents only choose from a limited set of short answers instead of generating an arbitrary length response, so far such tasks have been very close to the instruction following setting.

3.2 Language-assisted RL

In this section, we consider work that explores how knowledge about the structure of the world can be transferred from natural language corpora and methods into RL tasks, in cases where language itself is not essential to the task. Textual information can assist learning by specifying informative features, annotating states or entities in the environment, or describing subtasks in a multitask setting. In most cases covered here, the textual information is task-specific, with a few cases of using task-independent information through language parsers [Branavan *et al.*, 2012] and pre-trained sentence embeddings [Goyal *et al.*, 2019].

Language for Communicating Domain Knowledge

In a more general setting than instruction following, any kind of text containing potentially task-relevant information could be available. Such text may contain advice regarding the policy an agent should follow or information about the environment dynamics (see Figure 1). Unstructured and descriptive (in contrast to instructive) textual information is more abundant and can be found in wikis, manuals, books, or the web. However, using such information requires (i) retrieving useful information for a given context and (ii) grounding that information with respect to observations.

[Eisenstein *et al.*, 2009] learn abstractions in the form of conjunctions of predicate-argument structures that can reconstruct sentences and syntax in task-relevant documents using a generative language model. These abstractions are used to obtain a feature space that improves imitation learning outcomes. [Branavan *et al.*, 2012] learn a Q-function that

improves Monte-Carlo tree search planning for the first few moves in Civilization II, a turn-based strategy game, while accessing the game's natural language manual. Features for their Q-function depend on the game manual via learned sentence-relevance and predicate-labeling models whose parameters are optimised only by minimising the Q-function estimation error. Due to the structure of their hand-crafted features (some of which match states and actions to words, *e.g.* `action == irrigate AND action-word == "irrigate"`), these language processing models nonetheless somewhat learn to extract relevant sentences and classify their words as relating to the state, action, or neither. More recently, [Narasimhan *et al.*, 2018] investigate planning in a 2D game environment where properties of entities in the environment are annotated by natural language (*e.g.* the 'spider' and 'scorpion' entities might be annotated with the descriptions "randomly moving enemy" and "an enemy who chases you", respectively). Descriptive annotations facilitate transfer by learning a mapping between the annotations and the transition dynamics of the environment.

Language for Structuring Policies

One use of natural language is communicating information about the state and/or dynamics of an environment. As such it is an interesting candidate for constructing priors on the model structure or representations of an agent. This could include shaping representations towards more generalizable abstractions, making the representation space more interpretable to humans, or efficiently structuring the computations within a model.

[Andreas *et al.*, 2016] propose a neural architecture that is dynamically composed of a collection of jointly-trained neural modules, based on the parse tree of a natural language prompt. While originally developed for visual question answering, [Das *et al.*, 2018b] and [Bahdanau *et al.*, 2019] successfully apply variants of this idea to RL tasks. [Andreas *et al.*, 2018] explore the idea of natural language descriptions as a policy parametrization in a 2D navigation task adapted from [Janner *et al.*, 2018]. In a pre-training phase, the agent learns to imitate expert trajectories conditioned on instructions. By searching in the space of natural language instructions, the agent is then adapted to the new task where instructions and expert trajectories are not available.

The hierarchical structure of natural language and its compositionality make it a particularly good candidate for representing policies in hierarchical RL. [Andreas *et al.*, 2017] and [Shu *et al.*, 2018] can be viewed as using language (rather than logical or learned representations) as policy specifications for a hierarchical agent. More recently, [Hu *et al.*, 2019] consider generated natural language as a representation for macro actions in a real-time strategy game environment based on [Tian *et al.*, 2017]. In an IL setting, a meta-controller is trained to generate a sequence of natural language instructions. Simultaneously, a base-controller policy is trained to execute these generated instructions through a sequence of actions.

4 Trends for Natural Language in RL

The preceding sections surveyed the literature exploring how natural language can be integrated with RL. Several trends are evident: (i) studies for language-conditional RL are more numerous than for language-assisted RL, (ii) learning from task-dependent text is more common than learning from task-independent text, (iii) within work studying transfer from task-dependent text, only a handful of papers study how to use unstructured and descriptive text, (iv) there are only a few papers exploring methods for structuring internal plans and building compositional representations using the structure of language, and finally (v) natural language, as opposed to synthetically generated languages, is still not the standard in research on instruction following.

To advance the field, we argue that more research effort should be spent on learning from naturally occurring text corpora in contrast to instruction following. While learning from unstructured and descriptive text is particularly difficult, it has a much greater application range and potential for impact. Moreover, we argue for the development of more diverse environments with real-world semantics. The tasks used so far use small and synthetic language corpora and are too artificial to significantly benefit from transfer from real-world textual corpora. In addition, we emphasize the importance of developing standardized environments and evaluations for comparing and measuring progress of models that integrate natural language into RL agents.

We believe that there are several factors that make focusing such efforts worthwhile now: (i) recent progress in pre-training language models, (ii) general advances in representation learning, as well as (iii) development of tools that make constructing rich and challenging RL environments easier. Some significant work, especially in language-assisted RL, has been done prior to the surge of deep learning methods [Eisenstein *et al.*, 2009; Branavan *et al.*, 2012], and is worth revisiting. In addition, we encourage the reuse of software infrastructure, e.g. [Bahdanau *et al.*, 2019; Côté *et al.*, 2018; Chevalier-Boisvert *et al.*, 2019] for constructing environments and standardized tests.

4.1 Learning from Text Corpora in the Wild

The web contains abundant textual resources that provide instructions and how-to's.[2] For many games (which are often used as testbeds for RL), detailed walkthroughs and strategy guides exist. We believe that transfer from task-independent corpora could also enable agents to better utilize such task-dependent corpora. Preliminary results that demonstrate zero-shot capabilities [Radford *et al.*, 2019] suggest that a relatively small dataset of instructions or descriptions could suffice to ground and consequently utilize task-dependent information for better sample efficiency and generalization of RL agents.

Task-independent Corpora

Natural language reflects human knowledge about the world [Zellers *et al.*, 2018]. For instance, an effective language model should assign a higher probability to "get the green apple from the tree behind the house" than to "get the green tree from the apple behind the house". Harnessing such implicit commonsense knowledge captured by statistical language models could enable transfer of knowledge to RL agents.

In the short-term, we anticipate more use of pre-trained word and sentence representations for research on language-

---

[2]*e.g.* https://www.wikihow.com/ or https://stackexchange.com/

conditional RL. For example, consider instruction following with natural language annotations. Without transfer from a language model (or another language grounding task as in [Yu *et al.*, 2018]), instruction-following systems cannot generalize to instructions outside of the training distribution containing unseen synonyms or paraphrases (*e.g.* "fetch a stick", "return with a stick", "grab a stick and come back"). While pre-trained word and sentence representations alone will not solve the problem of grounding an unseen object or action, they do help with generalization to instructions with similar meaning but unseen words and phrases. In addition, we believe that learning representations for transferring knowledge about analogies, going beyond using analogies as auxiliary tasks [Oh *et al.*, 2017] will play an important role in generalizing to unseen instructions.

As pre-trained language models and automated question answering systems become more capable, one interesting long-term direction is studying agents that can query knowledge more explicitly. For example, during the process of planning in natural language, an agent that has a pre-trained language model as sub-component could let the latter complete "to open the door, I need to..." with "turn the handle". Such an approach could be expected to learn more rapidly than *tabula rasa* reinforcement learning. However, such agents would need to be capable of reasoning and planning in natural language, which is a related line of work (see Language for Structuring Policies in §3.2).

Task-dependent Corpora

Research on transfer from descriptive task-dependent corpora is promising due to its wide application potential. It also requires development of new environments, as early research may require access to relatively structured and partially grounded forms of descriptive language similarly to [Narasimhan *et al.*, 2018]. One avenue for early research is developing environments with relatively complex but still synthetic languages, providing information about environmental dynamics or advice about good strategies. For example, in works studying transfer from descriptive task-dependent language corpora [Narasimhan *et al.*, 2018], natural language sentences could be embedded using representations from pre-trained language models. Integrating and fine-tuning pre-trained information retrieval and machine reading systems similar to [Chen *et al.*, 2017a] with RL agents that query them could help in extracting and utilizing relevant information from unstructured task-specific language corpora such as game manuals as used in [Branavan *et al.*, 2012].

4.2 Towards Diverse Environments with Real-World Semantics

One of the central promises of language in RL is the ability to rapidly specify and help agents adapt to new goals, reward functions, and environment dynamics. This capability is not exercised at all by standard RL benchmarks like strategy games (which typically evaluate agents against a single or small number of fixed reward functions). It is evaluated in only a limited way by existing instruction following benchmarks, which operate in closed-task domains (navigation, object manipulation, etc.) and closed worlds. The simplicity of these tasks is often reflected in the simplicity of the language that describes them, with small vocabulary sizes and multiple pieces of independent evidence for the grounding of each word.

Real natural language has important statistical properties, such as the power-law distribution of word frequencies [Zipf, 1949] which does not appear in environments with synthetically generated language and small numbers of entities. Without environments that encourage humans to process (and force agents to learn from) complex composition and the "long tail" of lexicon entries, we cannot expect to hope that RL agents will generalize at all outside of *closed-world tasks*.

One starting point is provided by extending 3D house simulation environments [Gordon *et al.*, 2018; Das *et al.*, 2018a; Yan *et al.*, 2018] some of which already support generation and evaluation of templated instructions. These environments contain more real-world semantics (e.g. a microwave can be opened, a car is in the garage). However, interactions with the objects and the available templates for language instructions have been very limited so far. Another option is presented by open-world video games like Minecraft [Johnson *et al.*, 2016], in which users are free to assemble complex structures from simple parts, and thus have an essentially unlimited universe of possible objects to describe and goals involving those objects. More work is needed in exploring how learning grounding scales with the number of human annotations and environmental interactions [Bahdanau *et al.*, 2019; Chevalier-Boisvert *et al.*, 2019]. Looking ahead, as core machine learning tools for learning from feedback and demonstrations become sample-efficient enough to use in the real world, we anticipate that approaches combining language and RL will find applications as wide-ranging as autonomous vehicles, virtual assistants and household robots.

5 Conclusion

The currently predominant way RL agents are trained restricts their use to environments where all information about the policy can be gathered from directly acting in and receiving reward from the environment. This *tabula rasa* learning results in low sample efficiency and poor performance when transferring to other environments. Utilizing natural language in RL agents could drastically change this by transferring knowledge from natural language corpora to RL tasks, as well as between tasks, consequently unlocking RL for more diverse and real-world tasks. While there is a growing body of papers that incorporate language into RL, most of the research effort has been focused on simple RL tasks and synthetic languages, with highly structured and instructive text.

To realize the potential of language in RL, we advocate for more research into learning from unstructured or descriptive language corpora, with a greater use of NLP tools like pre-trained language models. Such research also requires development of more challenging environments that reflect the semantics and diversity of the real world.

Acknowledgements

References

[Agarwal *et al.*, 2019] Rishabh Agarwal, Chen Liang, Dale Schuurmans, and Mohammad Norouzi. Learning to generalize from sparse and underspecified rewards. *CoRR*, abs/1902.07198, 2019.

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

[Andreas and Klein, 2015] Jacob Andreas and Dan Klein. Alignment-based compositional semantics for instruction following. *ACL*, 2015.

[Andreas *et al.*, 2016] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *CVPR*, 2016.

[Andreas *et al.*, 2017] Jacob Andreas, Dan Klein, and Sergey Levine. Modular Multitask Reinforcement Learning with Policy Sketches. In *ICML*, 2017.

[Andreas *et al.*, 2018] Jacob Andreas, Dan Klein, and Sergey Levine. Learning with latent language. In *NAACL-HLT*, 2018.

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, 2015.

[Artzi and Zettlemoyer, 2013] Yoav Artzi and Luke Zettlemoyer. Weakly Supervised Learning of Semantic Parsers for Mapping Instructions to Actions. In *ACL*, 2013.

[Arulkumaran *et al.*, 2017] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. A Brief Survey of Deep Reinforcement Learning. *IEEE Signal Proc. Magazine*, 34(6):26–38, 2017.

[Babes *et al.*, 2011] Monica Babes, Vukosi Marivate, Kaushik Subramanian, and Michael L Littman. Apprenticeship learning about multiple intentions. In *ICML*, 2011.

[Bahdanau *et al.*, 2019] Dzmitry Bahdanau, Felix Hill, Jan Leike, Edward Hughes, Arian Hosseini, Pushmeet Kohli, and Edward Grefenstette. Learning to Understand Goal Specifications by Modelling Reward. In *ICLR*, 2019.

[Banko *et al.*, 2007] Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction from the web. In *IJCAI*, 2007.

[Barto and Mahadevan, 2003] Andrew G Barto and Sridhar Mahadevan. Recent advances in hierarchical reinforcement learning. *Discrete event dynamic systems*, 13(1-2):41–77, 2003.

[Bisk *et al.*, 2016] Yonatan Bisk, Deniz Yuret, and Daniel Marcu. Natural language communication with robots. In *ACL*, 2016.

[Bouziane *et al.*, 2015] Abdelghani Bouziane, Djelloul Bouchiha, Noureddine Doumi, and Mimoun Malki. Question answering systems: survey and trends. *Procedia Computer Science*, 73:366–375, 2015.

[Branavan *et al.*, 2010] S. R. K. Branavan, Luke S Zettlemoyer, and Regina Barzilay. Reading between the lines: Learning to map high-level instructions to commands. In *ACL*, 2010.

[Branavan *et al.*, 2012] S. R. K. Branavan, David Silver, and Regina Barzilay. Learning to Win by Reading Manuals in a Monte-Carlo Framework. *JAIR*, 2012.

[Chaplot *et al.*, 2018] Devendra Singh Chaplot, Kanthashree Mysore Sathyendra, Rama Kumar Pasumarthi, Dheeraj Rajagopal, and Ruslan Salakhutdinov. Gated-Attention Architectures for Task-Oriented Language Grounding. In *AAAI*, 2018.

[Chen and Mooney, 2011] David L Chen and Raymond J Mooney. Learning to interpret natural language navigation instructions from observations. In *AAAI*, 2011.

[Chen *et al.*, 2017a] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. In *ACL*, 2017.

[Chen *et al.*, 2017b] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter*, 19(2):25–35, 2017.

[Chen *et al.*, 2018] Howard Chen, Alane Suhr, Dipendra Kumar Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. *CoRR*, abs/1811.12354, 2018.

[Chevalier-Boisvert *et al.*, 2019] Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. BabyAI: A Platform to Study the Sample Efficiency of Grounded Language Learning. In *ICLR*, 2019.

[Côté *et al.*, 2018] Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew J. Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. Textworld: A learning environment for text-based games. *CoRR*, abs/1806.11532, 2018.

[Das *et al.*, 2018a] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied Question Answering. In *CVPR*, 2018.

[Das *et al.*, 2018b] Abhishek Das, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Neural Modular Control for Embodied Question Answering. *CoRL*, 2018.

[Deerwester *et al.*, 1990] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[DePristo and Zubek, 2001] Mark A DePristo and Robert Zubek. being-in-the-world. In *AAAI*, 2001.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[Dinan *et al.*, 2018] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. *CoRR*, abs/1811.01241, 2018.

[Eisenstein *et al.*, 2009] Jacob Eisenstein, James Clarke, Dan Goldwasser, and Dan Roth. Reading to learn: constructing features from semantic abstracts. In *ACL*, 2009.

[Firth, 1957] John R Firth. A synopsis of linguistic theory. *Special Volume of the Philological Society*, 1957.

[Frome *et al.*, 2013] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc Aurelio Ranzato, and Tomas Mikolov. DeViSE: A Deep Visual-Semantic Embedding Model. In *NIPS*, 2013.

[Fu *et al.*, 2019] Justin Fu, Anoop Korattikara, Sergey Levine, and Sergio Guadarrama. From Language to Goals: Inverse Reinforcement Learning for Vision-Based Instruction Following. In *ICLR*, 2019.

[Fulda *et al.*, 2017] Nancy Fulda, Daniel Ricks, Ben Murdoch, and David Wingate. What can you do with a rock? affordance extraction via word embeddings. *CoRR*, abs/1703.03429, 2017.

[Goldberg, 2019] Yoav Goldberg. Assessing BERT's Syntactic Abilities. *CoRR*, abs/1901.05287, 2019.

[Gopnik and Meltzoff, 1987] Alison Gopnik and Andrew Meltzoff. The development of categorization in the second year and its relation to other cognitive and linguistic developments. *Child development*, pages 1523–1531, 1987.

[Gordon *et al.*, 2018] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. Iqa: Visual question answering in interactive environments. In *CVPR*, 2018.

[Goyal *et al.*, 2019] Prasoon Goyal, Scott Niekum, and Raymond J. Mooney. Using Natural Language for Reward Shaping in Reinforcement Learning. *CoRR*, abs/1903.02020, 2019.

[Hermann *et al.*, 2017] Karl Moritz Hermann, Felix Hill, Simon Green, Fumin Wang, Ryan Faulkner, Hubert Soyer, David Szepesvari, Wojciech Marian Czarnecki, Max Jaderberg, Denis Teplyashin, Marcus Wainwright, Chris Apps, Demis Hassabis, and Phil Blunsom. Grounded language learning in a simulated 3d world. *CoRR*, abs/1706.06551, 2017.

[Ho and Ermon, 2016] Jonathan Ho and Stefano Ermon. Generative Adversarial Imitation Learning. In *NIPS*, 2016.

[Howard and Ruder, 2018] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *ACL*, 2018.

[Hu *et al.*, 2019] Hengyuan Hu, Denis Yarats, Qucheng Gong, Yuandong Tian, and Mike Lewis. Hierarchical decision making by generating and following natural language instructions. *arXiv preprint arXiv:1906.00744*, 2019.

[Infocom, 1980] Infocom. Zork I, 1980.

[Janner *et al.*, 2018] Michael Janner, Karthik Narasimhan, and Regina Barzilay. Representation learning for grounded spatial reasoning. *TACL*, 2018.

[Johnson *et al.*, 2016] Matthew Johnson, Katja Hofmann, Tim Hutton, and David Bignell. The malmo platform for artificial intelligence experimentation. In *IJCAI*, 2016.

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

[Kollar *et al.*, 2010] Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. Toward understanding natural language directions. In *HRI*, 2010.

[Kostka *et al.*, 2017] B. Kostka, J. Kwiecieli, J. Kowalski, and P. Rychlikowski. Text-based adventures of the golovin AI agent. In *2017 IEEE Conference on Computational Intelligence and Games*, pages 181–188, 2017.

[Kuhlmann *et al.*, 2004] Gregory Kuhlmann, Peter Stone, Raymond Mooney, and Jude Shavlik. Guiding a Reinforcement Learner with Natural Language Advice: Initial Results in RoboCup Soccer. In *AAAI-2004 workshop on supervisory control of learning and adaptive systems*, 2004.

[MacGlashan *et al.*, 2015] James MacGlashan, Monica Babes-Vroman, Marie desJardins, Michael L. Littman, Smaranda Muresan, Shawn Squire, Stefanie Tellex, Dilip Arumugam, and Lei Yang. Grounding english commands to reward functions. In *Robotics: Science and Systems XI*, 2015.

[MacMahon *et al.*, 2006] Matt MacMahon, Brian Stankiewicz, and Benjamin Kuipers. Walk the talk: Connecting language, knowledge, and action in route instructions. In *AAAI*, 2006.

[Massiceti *et al.*, 2018] Daniela Massiceti, N Siddharth, Puneet K Dokania, and Philip HS Torr. Flipdial: A generative model for two-way visual dialogue. In *CVPR*, 2018.

[Mei *et al.*, 2016] Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. Listen, Attend, and Walk: Neural Mapping of Navigational Instructions to Action Sequences. *AAAI*, 2016.

[Mey, 1993] J. Mey. *Pragmatics: An Introduction*. Blackwell, 1993.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*, 2013.

[Misra *et al.*, 2017] Dipendra Misra, John Langford, and Yoav Artzi. Mapping Instructions and Visual Observations to Actions with Reinforcement Learning. *EMNLP*, 2017.

[Narasimhan *et al.*, 2015] Karthik Narasimhan, Tejas D. Kulkarni, and Regina Barzilay. Language understanding for text-based games using deep reinforcement learning. In *EMNLP*, 2015.

[Narasimhan *et al.*, 2018] Karthik Narasimhan, Regina Barzilay, and Tommi Jaakkola. Grounding Language for Transfer in Deep Reinforcement Learning. *JAIR*, 2018.

[Oh *et al.*, 2017] Junhyuk Oh, Satinder P. Singh, Honglak Lee, and Pushmeet Kohli. Zero-shot task generalization with multi-task deep reinforcement learning. In *ICML*, 2017.

[Osa *et al.*, 2018] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, Jan Peters, et al. An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics*, 7(1-2):1–179, 2018.

[Peters *et al.*, 2018a] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL*, 2018.

[Peters *et al.*, 2018b] Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. Dissecting contextual word embeddings: Architecture and representation. In *EMNLP*, 2018.

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

[Shu *et al.*, 2018] Tianmin Shu, Caiming Xiong, and Richard Socher. Hierarchical and Interpretable Skill Acquisition in Multi-task Reinforcement Learning. *ICLR*, 2018.

[Shusterman *et al.*, 2011] Anna Shusterman, Sang Ah Lee, and Elizabeth Spelke. Cognitive effects of language on human navigation. *Cognition*, 120(2):186–201, 2011.

[Silver *et al.*, 2017] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354, 2017.

[Singh *et al.*, 2002] Satinder Singh, Diane Litman, Michael Kearns, and Marilyn Walker. Optimizing dialogue management with reinforcement learning: Experiments with the njfun system. *JAIR*, 2002.

[Socher *et al.*, 2013] Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. Zero-shot Learning Through Cross-modal Transfer. In *NIPS*, 2013.

[Spelke and Kinzler, 2007] Elizabeth Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007.

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

[Tellex *et al.*, 2011] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI*, 2011.

[Tenney *et al.*, 2019] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. In *ICLR*, 2019.

[Tesauro, 1995] Gerald Tesauro. Temporal Difference Learning and TD-Gammon. *Communications of the ACM*, 38(3):58–69, 1995.

[Tian *et al.*, 2017] Yuandong Tian, Qucheng Gong, Wenling Shang, Yuxin Wu, and C. Lawrence Zitnick. ELF: An Extensive, Lightweight and Flexible Research Platform for Real-time Strategy Games. *NIPS*, 2017.

[Torrado *et al.*, 2018] Ruben Rodriguez Torrado, Philip Bontrager, Julian Togelius, Jialin Liu, and Diego Perez-Liebana. Deep reinforcement learning for general video game ai. In *2018 IEEE Conference on Computational Intelligence and Games*, pages 1–8, 2018.

[Tsividis *et al.*, 2017] Pedro Tsividis, Thomas Pouncy, Jaqueline L. Xu, Joshua B. Tenenbaum, and Samuel J. Gershman. Human learning in atari. In *AAAI*, 2017.

[Wang *et al.*, 2016] Sida I Wang, Percy Liang, and Christopher D Manning. Learning Language Games through Interaction. In *ACL*, 2016.

[Wang *et al.*, 2018] Xin Wang, Qiuyuan Huang, Asli Çelikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. *CoRR*, abs/1811.10092, 2018.

[White and Sofge, 1992] David Ashley White and Donald A Sofge. *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptative Approaches*. Van Nostrand Reinhold Company, 1992.

[Yan *et al.*, 2018] Claudia Yan, Dipendra Kumar Misra, Andrew Bennett, Aaron Walsman, Yonatan Bisk, and Yoav Artzi. CHALET: cornell house agent learning environment. *CoRR*, abs/1801.07357, 2018.

[Yu *et al.*, 2018] Haonan Yu, Haichao Zhang, and Wei Xu. Interactive Grounded Language Acquisition and Generalization in a 2d World. *ICLR*, 2018.

[Yuan *et al.*, 2018] Xingdi Yuan, Marc-Alexandre Côté, Alessandro Sordoni, Romain Laroche, Remi Tachet des Combes, Matthew J. Hausknecht, and Adam Trischler. Counting to explore and generalize in text-based games. *CoRR*, abs/1806.11525, 2018.

[Zellers *et al.*, 2018] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *EMNLP*, 2018.

[Ziebart *et al.*, 2008] Brian D Ziebart, Andrew Maas, J Andrew Bagnell, and Anind K Dey. Maximum Entropy Inverse Reinforcement Learning. In *AAAI*, 2008.

[Zipf, 1949] George Kingsley Zipf. Human behavior and the principle of least effort. 1949.