Experience Grounds Language

Yonatan Bisk* **Ari Holtzman*** **Jesse Thomason***

Jacob Andreas Yoshua Bengio Joyce Chai Mirella Lapata
Angeliki Lazaridou Jonathan May Aleksandr Nisnevich Nicolas Pinto Joseph Turian

arXiv:2004.10151v1 [cs.CL] 21 Apr 2020

Abstract

Successful linguistic communication relies on a shared experience of the world, and it is this shared experience that makes utterances meaningful. Despite the incredible effectiveness of language processing models trained on text alone, today's best systems still make mistakes that arise from a failure to relate language to the physical world it describes and to the social interactions it facilitates.

Natural Language Processing is a diverse field, and progress throughout its development has come from new representational theories, modeling techniques, data collection paradigms, and tasks. We posit that the present success of representation learning approaches trained on large text corpora can be deeply enriched from the parallel tradition of research on the contextual and social nature of language.

In this article, we consider work on the contextual foundations of language: grounding, embodiment, and social interaction. We describe a brief history and possible progression of how contextual information can factor into our representations, with an eye towards how this integration can move the field forward and where it is currently being pioneered. We believe this framing will serve as a roadmap for truly contextual language understanding.

Improvements in hardware and data collection have galvanized progress in NLP. Performance peaks have been reached in tasks such as language modeling (Radford et al., 2019; Zellers et al., 2019c; Keskar et al., 2019) and span-selection question answering (Devlin et al., 2019a; Yang et al., 2019; Lan et al., 2019) through massive data and massive models. Now is an excellent time to reflect on the direction of our field, and on the relationship of language to the broader AI, Cognitive Science, and Linguistics communities.

We are interested in how the data and world a language learner is exposed to defines and potentially constrains the scope of the learner's semantics. Meaning is not a unique property of language; it is a universal property of intelligence. Consequently, we must consider what knowledge or concepts are missing from models trained solely on text corpora, even when those corpora are large scale or meticulously annotated.

> Meaning is not a unique property of language, but a general characteristic of human activity ... We cannot say that each morpheme or word has a single or central meaning, or even that it has a continuous or coherent range of meanings ... there are two separate uses and meanings of language – the concrete ... and the abstract.
>
> *Zellig S. Harris (Distributional Structure 1954)*

Large, generative language models emit utterances which violate visual, physical, and social commonsense. Perhaps this is to be expected, since they are tested in terms of Independent and Identically Distributed held-out sets, rather than queried on points meant to probe the granularity of the distinctions they make (Kaushik et al., 2020; Gardner et al., 2020). We draw on previous work in NLP, Cognitive Science, and Linguistics to provide a roadmap towards addressing these gaps. We posit that the universes of knowledge and experience available to NLP models can be defined by successively larger world scopes: from a single corpus to a fully embodied and social context.

We propose the notion of a World Scope (WS) as a lens through which to view progress in NLP. From the beginning of corpus linguistics (Zipf, 1932; Harris, 1954), to the formation of the Penn Treebank (Marcus et al., 1993), NLP researchers have consistently recognized the limitations of corpora in terms of coverage of language *and* experience. To organize this intuition, we propose five WSs, and note that most current work in NLP operates in the second (internet-scale data).

We define five levels of **World Scope**:

WS1. Corpus *(our past)*
WS2. Internet *(our present)*
WS3. Perception
WS4. Embodiment
WS5. Social

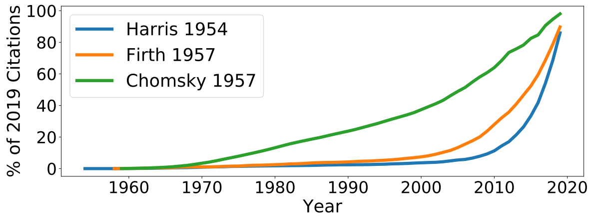
1 WS1: Corpora and Representations

The story of computer-aided, data-driven language research begins with the corpus. While by no means the only of its kind, the Penn Treebank (Marcus et al., 1993) is the canonical example of a sterilized subset of naturally generated language, processed and annotated for the purpose of studying representations, a perfect example of WS1. While initially much energy was directed at finding *structure* (e.g. syntax trees), this has been largely overshadowed by the runaway success of unstructured, fuzzy representations—from dense word vectors (Mikolov et al., 2013) to contextualized pretrained representations (Peters et al., 2018b; Devlin et al., 2019b).

Yet fuzzy, conceptual word representations have a long history that predates the recent success of deep learning methods. Philosophy (Lakoff, 1973) and linguistics (Coleman and Kay, 1981) recognized that meaning is flexible yet structured. Early experiments on neural networks trained with sequences of words (Elman, 1990) suggested that vector representations could capture both syntax and semantics. Subsequent experiments with larger models, document contexts, and corpora have demonstrated that representations learned from text capture a great deal of information about word meanings in and out of context (Bengio et al., 2003; Collobert and Weston, 2008; Turian et al., 2010; Mikolov et al., 2013; McCann et al., 2017)

It has long been acknowledged that context lends meaning (Firth, 1957; Turney and Pantel, 2010). Local contexts proved powerful when combined with agglomerative clustering guided by mutual information (Brown et al., 1992). A word's position in this hierarchy captures semantic and syntactic distinctions. When the Baum-Welch algorithm (Welch, 2003) was applied to unsupervised Hidden Markov Models, it assigned a class distribution to every word, and that distribution was considered a partial representation of a word's "meaning." If the set of classes was small, syntax-like classes were induced; If the set was large, classes became more semantic. Over the years this "search for meaning" has played out again and again, often with similar themes.

Later approaches discarded structure in favor of larger context windows for better semantics. The most popular generative approach came in the form of Latent Dirichlet Allocation (Blei et al., 2003). The sentence structure was discarded and a document generated as a bag-of-words conditioned on topics. However, the most successful vector representations came from Latent Semantic Indexing/Analysis (Deerwester et al., 1988, 1990; Dumais, 2004), which represents words by their co-occurrence with other words. Via singular value decomposition, these matrices are reduced to low dimensional projections, essentially a bag-of-words.



Academic interest in Firth and Harris increases dramatically around 2010, perhaps due to the popularization of Firth (1957) "You shall know a word by the company it keeps."

A related question also existed in connectionism (Pollack, 1987; James and Miikkulainen, 1995): Are concepts distributed through edges or local to units in an artificial neural network?

> *"... there has been a long and unresolved debate between those who favor localist representations in which each processing element corresponds to a meaningful concept and those who favor distributed representations."* Hinton (1990)
> *Special Issue on Connectionist Symbol Processing*

Unlike work that defined words as distributions over clusters, which were perceived as having intrinsic meaning to the user of a system, in connectionism there was the question of where *meaning* resides. The tension of modeling symbols and distributed representations was nicely articulated by Smolensky (1990), and alternative representations (Kohonen, 1984; Hinton et al., 1986; Barlow, 1989) and approaches to structure and composition (Erk and Padó, 2008; Socher et al., 2012) span decades of research.

All of these works rely on corpora. The Brown Corpus (Francis, 1964) and Penn Treebank (Marcus et al., 1993) were colossal undertakings that defined linguistic context for decades. Only relatively recently (Baroni et al., 2009) has the cost of annotations decreased enough to enable the introduction of new tasks and have web-crawls become viable for researchers to collect and process (WS2).[1]

[1] An important parallel discussion centers on the hardware

2 WS2: The Written World

Corpora in NLP have broadened to include large web-crawls. The use of unstructured, unlabeled, multi-domain, and multilingual data broadens our world scope, in the limit, to everything humanity has ever written. We are no longer constrained to a single author or source, and the temptation for NLP is to believe everything that needs knowing can be learned from the written world.

This move towards using large scale (whether mono- or multilingual) has led to substantial advances in performance on existing and novel community benchmarks (Wang et al., 2019a). These advances have come due to the transfer learning enabled by representations in deep models. Traditionally, transfer learning relied on our understanding of model classes, such as English grammar. Domain adaptation could proceed by simply providing a model with sufficient data to capture lexical variation, while assuming most higher-level structure would remain the same. Embeddings—lexical representations built from massive corpora—encompass multiple domains and lexical senses, violating this structural assumption.

These representations require scale—both in terms of data and parameters. Concretely, Mikolov et al. (2013) trained on 1.6 billion tokens, while Pennington et al. (2014) scaled up to 840 billion tokens from Common Crawl. Recent approaches have made progress by substantially increasing the number of model parameters to better consume these vast quantities of data. Peters et al. (2018a) introduced ELMo, containing roughly 94M parameters, Devlin et al. (2019b) introduced BERT-Large with 340M, and Radford et al. (2019); Zellers et al. (2019c) experimented with 1.5B parameter Transformers (Vaswani et al., 2017) trained on over 120GB of text.

Our interest in these approaches is two-fold:

1. Larger models see diminishing returns (especially with respect to computational cost) despite the massive availability of data.

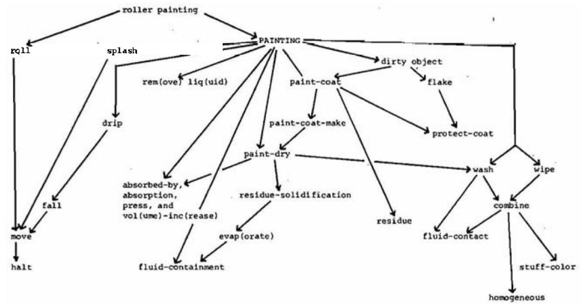
2. These models have expanded the notion of local context to include multiple sentences.

Current models are the next (impressive) step in modeling lexical distributions which started with Good (1953), the weights of Kneser and Ney (1995); Chen and Goodman (1996), and the power-law distributions of Teh (2006). Modern approaches to learning dense representations allow us to better estimate these distributions from massive corpora. However, modeling lexical co-occurrence, no matter the scale, is still "modeling the *written* world." Models constructed this way blindly search for symbolic co-occurences void of meaning.

In regards to the apparent paradox of "impressive results" vs. "diminishing returns", language modeling—the modern workhorse of neural NLP systems—provides a poignant example. Recent pretraining literature has produced results that few could have predicted, crowding leader boards results that make claims to super-human accuracy (Rajpurkar et al., 2018). However, there are diminishing returns. For example, in the LAMBADA dataset (Paperno et al., 2016) (designed to capture human intuition), GPT2 (Radford et al., 2019) (1.5B), Megatron-LM (Shoeybi et al., 2019) (8.3B), and TuringNLG (Rosset, 2020) (17B) perform within a few points of each other and very far from perfect (<68%). We argue that continuing to expand hardware and data sizes by orders of magnitude is not the path forward.

The aforementioned approaches for learning transferable representations demonstrate that sentence and document context provide powerful signals for learning aspects of meaning, especially semantic relations among words (Fu et al., 2014) and inferential relationships among sentences (Wang et al., 2018). The extent to which they capture deeper notions of contextual meaning, however, remains an open question. On the one hand, state-of-the-art, pre-trained language models are capable of generating locally coherent narratives (Radford et al., 2019). On the other hand, past work has found that pretrained word and sentence representations fail to capture many grounded features of words (Lucy and Gauthier, 2017) and sentences, while current NLU systems fail on the thick tail of experience-informed inferences, such as hard coreference problems (Peng et al., 2015). As pretraining schemes seem to be reaching the point of diminishing returns from data, even for some syntactic phenomena (van Schijndel et al., 2019), we posit that other forms of supervision, such as multimodal perception, will be necessary to learn the remaining aspects of meaning in context.

required to enable advances that move us from one world scope to the next. Playstations (Pinto et al., 2009) and GPUs (Krizhevsky et al., 2012) made much of WS2 advances possible, but perception, interaction, and social robotics bring new challenges which our current tools are not equipped to handle.



Eugene Charniak (A Framed PAINTING: The Representation of a Common Sense Knowledge Fragment 1977)

3 WS3: The World of Sights and Sounds

Language learning needs perception. An agent observing the real world can build axioms from which to extrapolate. Learned, physical heuristics, such as that a falling cat will land quietly, are generalized and abstracted into language metaphors like *as nimble as a cat* (Lakoff, 1980). World knowledge forms the basis for how people make entailment and reasoning decisions. There exists strong evidence that children require grounded sensory input, not just speech, to learn language (Sachs et al., 1981; O'Grady, 2005; Vigliocco et al., 2014).

Perception includes auditory, tactile, and visual input. Even restricted to purely linguistic signals, auditory input is necessary for understanding sarcasm, stress, and meaning implied through prosody. Further, tactile senses give meaning, both physical (Sinapov et al., 2014; Thomason et al., 2016) and abstract, to concepts like *heavy* and *soft*. Visual perception is a powerful signal for modeling a vastness of experiences in the world that cannot be documented by text alone (Harnad, 1990).

For example, frames and scripts (Schank and Abelson, 1977; Charniak, 1977; Dejong, 1981; Mooney and Dejong, 1985) require understanding often unstated sets of pre- and post-conditions about the world. To borrow from Charniak (1977), how should we learn the meaning, method, and implications of *painting*? A web crawl of knowledge from an exponential number of possible how-to, text-only guides and manuals (Bisk et al., 2020) is misdirected without *some* fundamental referents to ground symbols to. We posit that models must be able to watch and recognize objects, people, and activities to understand the language describing them (Li et al., 2019b; Krishna et al., 2017; Yatskar et al., 2016; Perlis, 2016).

While the NLP community has played an important role in the history of scripts and grounding (Mooney, 2008), recently remarkable progress in grounding has taken place in the Computer Vision community. In the last decade, CV researchers have refined and codified the "backbone" of computer vision architectures. These advances have lead to parameter efficient models (Tan and Le, 2019) and real time object detection on embedded devices and phones (Rastegari et al., 2016; Redmon and Farhadi, 2018).

We find anecdotally that many researchers in the NLP community still assume that vision models tested on the 1,000 class[2] ImageNet challenge are limited to extracting a bag of visual words. In reality, these architectures are mature and capable of generalizing, both from the perspective of engineering infrastructure[3] and that the backbones have been tested against other technologies like self-attention (Ramachandran et al., 2019). The stability of these architectures allows for new research into more challenging world modeling. Mottaghi et al. (2016) predict the effects of forces on objects in images. Bakhtin et al. (2019) extends this physical reasoning to complex puzzles of cause and effect. Sun et al. (2019b,a) model scripts and actions. Alternative, unsupervised training regimes (Bachman et al., 2019) also open up research towards automatic concept formation.

While a minority of language researchers make forays into computer vision, researchers in CV are often willing to incorporate language signals. Advances in computer vision architectures and modeling have enabled building semantic representations rich enough to interact with natural language. In the last decade of work descendant from image captioning (Farhadi et al., 2010; Mitchell et al., 2012), a myriad of tasks on visual question answering (Antol et al., 2015; Das et al., 2018), natural language and visual reasoning (Suhr et al., 2019b), visual commonsense (Zellers et al., 2019a), and multilingual captioning and translation via video (Wang et al., 2019b) have emerged. These combined text and vision benchmarks are rich enough to train large-scale, multimodal transformers (Li et al., 2019a; Lu et al., 2019; Zhou et al., 2019) without language pretraining, for example via conceptual captions (Sharma et al., 2018), or further broadened to include audio (Tsai et al., 2019).

[2]Or the 1,600 classes of Anderson et al. (2017).

[3]Torchvision and Detectron both include pretrained versions of over a dozen models.

Semantic level representations emerge from ImageNet classification pretraining partially due to class hypernyms. For example, the *person* class sub-divides into many professions and hobbies, like *firefighter*, *gymnast*, and *doctor*. To differentiate such sibling classes, learned vectors can also encode lower-level characteristics like clothing, hair, and typical surrounding scenes. These representations allow for pixel level masks and skeletal modeling, and can be extended to zero-shot settings targeting all 20,000 ImageNet categories (Chao et al., 2016; Changpinyo et al., 2017). Modern architectures are flexible enough to learn to differentiate instances within a general class, such as *face*. For example, facial recognition benchmarks require distinguishing over 10,000 unique faces (Liu et al., 2015). While vision is by no means "solved," vision benchmarks have led to off-the-shelf tools for building representations rich enough for tens of thousands of objects, scenes, and individuals.

That the combination of language and vision supervision leads to models that produce clear and coherent sentences, or that their attention can translate parts of an image to words, indicates they truly are learning about language. Moving forward, we believe benchmarks incorporating auditory, tactile, and visual sensory information together with language will be crucial. Such benchmarks will facilitate modeling language meaning with respect to an experienced *world*.

4 WS4: Embodiment and Action

> If A and B have some environments in common and some not ... we say that they have different meanings, the amount of meaning difference corresponding roughly to the amount of difference in their environments ...
>
> *Zellig S. Harris (Distributional Structure 1954)*

Many of the phenomena in Winograd (1972)'s *Understanding Natural Language* require representation of concepts derivable from perception, such as color, counting, size, shape, spatial reasoning, and stability. Recent work (Gardner et al., 2020) argues that these same dimensions can serve as meaningful perturbations of inputs to evaluate models' reasoning consistency. However, their presence is in service of actual interactions with the world. Action taking is a natural next rung on the ladder to broader context.

An embodied agent, whether in a virtual world, such as a 2D Maze (MacMahon et al., 2006), a Baby AI in a grid world (Chevalier-Boisvert et al., 2019), Vision-and-Language Navigation (Anderson et al., 2018; Thomason et al., 2019b), or the real world (Tellex et al., 2011; Matuszek, 2018; Thomason et al., 2020; Tellex et al., 2020) must translate from language to control. Control and action taking open several new dimensions to understanding and actively learning about the world. Queries can be resolved via dialog-based exploration with a human interlocutor (Liu and Chai, 2015), even as new object properties like texture, weight, and sound become available (Thomason et al., 2017). We see the need for embodied language with complex meaning when thinking deeply about even the most innocuous of questions:

> *Is an orange more like a baseball or a banana?*

WS1 is likely not to have an answer beyond that the objects are common nouns that can both be held. WS2 may also capture that oranges and baseballs both roll, but is unlikely to capture the deformation strength, surface texture, or relative sizes of these objects (Elazar et al., 2019). WS3 can begin to understand the relative deformability of these objects, but is likely to confuse how much force is necessary given that baseballs are used much more roughly than oranges in widely distributed media. WS4 can appreciate the nuanced nature of the question—the orange and baseball share a similar texture and weight allowing for similar manipulation, while the orange and banana both contain peels, deform under stress, and are edible. We as humans have rich representations of these items. The words evoke a myriad of properties, and that richness allows us to reason.

Control is where people first learn abstraction and simple examples of post-conditions through trial and error. The most basic scripts humans learn start with moving our own bodies and achieving simple goals as children, such as stacking blocks. In this space, we have unlimited supervision from the environment and can learn to generalize across plans and actions. In general, simple worlds do not entail simple concepts: even in block worlds (Bisk et al., 2018) where concepts like "mirroring"—one of a massive set of physical phenomena that we generalize and apply to abstract concepts—appear.

In addition to learning basic physical properties of the world from interaction, WS4 also allows the agent to construct rich pre-linguistic representations from which to generalize. Hespos and Spelke (2004) show pre-linguistic category formation within children that are then later codified by social constructs. Mounting evidence seems to indicate that children have trouble transferring knowledge from the 2D world of books (Barr, 2013) and iPads (Lin et al., 2017) to the physical 3D world. So while we might choose to believe that we can encode parameters (Chomsky, 1981) more effectively and efficiently than evolution provided us, developmental experiments indicate doing so without 3D interaction may prove insurmountable.

Part of the problem is that much of the knowledge humans hold about the world is intuitive, possibly incommunicable by language, but is required to understand that language. This intuitive knowledge could be acquired by embodied agents interacting with their environment, even before language words are grounded to meanings.

Robotics and embodiment are not available in the same off-the-shelf manner as computer vision models. However, there is rapid progress in simulators and commercial robotics, and as language researchers we should match these advances at every step. As action spaces grow, we can study complex language instructions in simulated homes (Shridhar et al., 2019) or map language to physical robot control (Blukis et al., 2019; Chai et al., 2018). The last few years have seen massive advances in both high fidelity simulators for robotics (Todorov et al., 2012; Coumans and Bai, 2016–2019; NVIDIA, 2019; Xiang et al., 2020) and the cost and availability of commodity hardware (Fitzgerald, 2013; Campeau-Lecours et al., 2019; Murali et al., 2019).

As computers transition from desktops to pervasive mobile and edge devices, we must make and meet the expectation that NLP can be deployed in any of these contexts. Current representations have very limited utility in even the most basic robotic settings (Scalise et al., 2019), making collaborative robotics (Rosenthal et al., 2010) largely a domain of custom engineering rather than science.

5 WS5: The Social World

> In order to talk about concepts, we must understand the importance of mental models... we set up a model of the world which serves as a framework in which to organize our thoughts. We abstract the presence of particular objects, having properties, and entering into events and relationships.
>
> *Terry Winograd - 1971*

Natural language arose to enable interpersonal communication (Dunbar, 1993). Since then, it has been used in everything from laundry instructions to personal diaries, which has given us data and situated scenarios that are the raw material for the previous levels of contextualization. Interpersonal communication in service of real-world cooperation is the prototypical use of language, and the ability to facilitate such cooperation remains the final test of a learned agent.

The acknowledgement that interpersonal communication is *the* necessary property of a linguistic intelligence is older than the terms "Computational Linguistics" or "Artificial Intelligence." Launched into computing when Alan Turing suggested the "Imitation Game" (Turing, 1950), the first illustrative example of the test brings to bear an issue that has haunted generative NLP since it was possible. Smoke and mirrors cleverness, in which situations are framed to the advantage of the model, can create the appearance of genuine content where there is none. This phenomenon has been noted countless times, from Pierce (1969) criticizing Speech Recognition as "deceit and glamour" (Pierce, 1969) to Marcus and Davis (2019) which complains of humanity's "gullibility gap".

Interpersonal dialogue is canonically framed as a grand test for AI (Norvig and Russel, 2002), but there are few to no examples of artificial agents one could imagine socializing with in anything more than transactional (Bordes et al., 2016) or extremely limited game scenarios (Lewis et al., 2017)—at least not ones that aren't purposefully hollow and fixed, where people can project whatever they wish (e.g. ELIZA (Weizenbaum, 1966)). More important to our discussion is why socialization is required as the next rung on the context ladder in order to fully ground meaning.

Language that Does Something Work in the philosophy of language has long suggested that *function* is the source of meaning, as famously illustrated through Wittgenstein's "language games" (Wittgenstein, 1953, 1958) and J.L. Austin's "performative speech acts" (Austin, 1975). That function is the source of meaning was echoed in linguistics usage-based theory of language acquisition, which suggests that constructions that are useful are the building blocks for everything else (Langacker, 1987, 1991). In recent years, this theory has begun to shed light on what use-cases language presents in both acquisition and its initial origins in our species (Tomasello, 2009; Barsalou, 2008).

WS1, WS2, WS3, and WS4 lend an extra depth to the interpretation of language through context, because they expand the factorizations of information available to define meaning. Understanding that what one says *can change what people do* allows language to take on its most active role. This is the ultimate goal for natural language generation: language that *does* something to the world.

Despite the current, passive way generation is created and evaluated (Liu et al., 2016), the ability to change the world through other people is the rich learning signal that natural language generation offers. In order to learn the effects language has on the world, an agent must *participate* in linguistic activity, such as negotiation (Lewis et al., 2017; He et al., 2018), collaboration (Chai et al., 2017), visual disambiguation (Liu and Chai, 2015; Anderson et al., 2018; Lazaridou et al., 2017), providing emotional support (Rashkin et al., 2019), all of which require inferring mental states and social outcomes—a key area of interest in itself (Zadeh et al., 2019).

As an example, what "hate" means in terms of discriminative information is always at question: it can be defined as "strong dislike," but what it tells one about the processes operating in the environment requires social context to determine (Bloom, 2002). It is the toddler's social experimentation with "I hate you" that gives the word weight and definite intent (Ornaghi et al., 2011). In other words, the discriminative signal for the most foundational part of a word's meaning can only be observed by its effect on the world, and active experimentation seems to be key to learning that effect. This is in stark contrast to the disembodied chat bots that are the focus of the current dialogue community (Adiwardana et al., 2020; Zhou et al., 2020; Chen et al., 2018; Serban et al., 2017), which often do not learn from individual experiences and are not given enough of a persistent environment to learn about the effects of actions.

Theory of Mind By attempting to get what we want, we confront people with their own desires and identities. Premack and Woodruff (1978) began to formalize how much the ability to consider the feelings and knowledge of others is human-specific and how it actually functions, commonly referred to now as the "Theory of Mind."

Q: Please write me a sonnet on the subject of the Forth Bridge.
A: Count me out on this one. I never could write poetry.
Q: Add 34957 to 70764.
A: (Pause about 30 seconds and then give as answer) 105621.
Q: Do you play chess?
A: Yes.
Q: I have K at my K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What do you play?
A: (After a pause of 15 seconds) R-R8 mate.

A.M. Turing (Computing Machinery and Intelligence 1950)

In the language community, this paradigm has been described under the "Speaker-Listener" model (Stephens et al., 2010), and a rich theory to describe this computationally is being actively developed under the Rational Speech Act Model (Frank and Goodman, 2012; Bergen et al., 2016).

Recently, a series of challenges that attempt to address this fundamental aspect of communication (Nematzadeh et al., 2018; Sap et al., 2019) have been introduced. Using traditional-style datasets can be problematic due to the risk of embedding spurious patterns (Le et al., 2019). Despite increased scores on datasets due to larger models and more complex pretraining curricula, it seems unlikely that models understand their listener any more than they understand the physical world in which they exist. This disconnect is driven home by studies of bias (Gururangan et al., 2018; Glockner et al., 2018) and techniques like Adversarial Filtering (Zellers et al., 2018, 2019b), which elucidate the biases models exploit in lieu of understanding.

Our training data, complex and large as it is, does not offer the discriminatory signals that make the hypothesizing of consistent identity or mental states an efficient path towards lowering perplexity or raising accuracy. Firstly, there is a lack of inductive bias (Martin et al., 2018): it is not clear that *any* learned model, not just a neural network, from reading alone would be able to posit that people exist, that they arrange narratives along an abstract single-dimension called *time*, and that they describe things in terms of causality. Models learn what they need to discriminate, and so inductive bias is not just about efficiency, it is about making sure what is learned will generalize (Mitchell, 1980). Secondly, current cross entropy training losses actively discourage learning the tail of the distribution properly, as events that are statistically infrequent are drowned out (Pennington et al., 2014; Holtzman et al., 2020). Researchers have just begun to scratch the surface by considering more dynamic evaluations (Zellers et al., 2020; Dinan et al., 2019), but true persistence of identity and adaption to change are still a long way off.

Language in a Social Context Whenever language is used between two people, it exists in a concrete social context: status, role, intention, and countless other variables intersect at a specific point (Wardhaugh, 2011). Currently, these complexities are overlooked by selecting labels crowd workers agree on. That is, our notion of ground truth in dataset construction is based on crowd consensus bereft of social context (Zellers et al., 2020).

In a real-world interaction, even one as anonymous as interacting with a cashier at a store, notions of social cognition (Fiske and Taylor, 1991; Carpenter et al., 1998) are present in the style of utterances and in the social script of the exchange. The true evaluation of generative models will require the construction of situations where artificial agents are considered to have enough identity to be granted *social standing* for these interactions.

Social interaction is a precious signal and some NLP researchers have began scratching the surface of it, especially for persuasion related tasks where interpersonal relationships are key (Wang et al., 2019c; Tan et al., 2016). These initial studies have been strained by the training-validation-test set scenario, as collecting static of this data is difficult for most use-cases and often impossible as natural situations cannot be properly facilitated. Learning by participation is a necessary step to the ultimately *social* venture of communication. By exhibiting different attributes and sending varying signals, the sociolinguistic construction of identity (Ochs, 1993) could be examined more deeply than ever before, yielding an understanding of social intelligence that simply isn't possible with a fixed corpus. This social layer is the foundation upon which language situates (Tomasello, 2009). Understanding this social layer is not only necessary, but will make clear complexities around implicature and commonsense that so obviously arise in corpora, but with such sparsity that they cannot properly be learned (Gordon et al., 2018). Once models are commonly expected to be interacted with in order to be tested, probing their decision boundaries for simplifications of reality as in Kaushik et al. will become trivial and natural.

6 Self-Evaluation

You can't learn language from the radio.

Nearly every NLP course will at some point make this claim. While the futility of learning language from an ungrounded linguistic signal is intuitive, those NLP courses go on to attempt precisely that learning. In fact, our entire field follows this pattern: trying to learn language from the *internet*, standing in as the modern radio to deliver limitless language. The need for language to attach to "extralinguistic events" (Ervin-Tripp, 1973) and the requirement for social context (Baldwin et al., 1996) should guide our research.

Concretely, we use our notion of World Scopes to make the following concrete claims:

You can't learn language ...

... from the radio (internet). **WS2 $\subset$ WS3**

A learner cannot be said to be in WS3 *if it can perform its task without sensory perception such as visual, auditory, or tactile information.*

... from a television. **WS3 $\subset$ WS4**

A learner cannot be said to be in WS4 *if the space of actions and consequences of its environment can be enumerated.*

... by yourself. **WS4 $\subset$ WS5**

A learner cannot be said to be in WS5 *if its cooperators can be replaced with cleverly pre-programmed agents to achieve the same goals.*

By these definitions, most of NLP research still resides in WS2, and we believe this is at odds with the goals and origins of our science. This does not invalidate the utility or need for any of the research within NLP, but it is to say that much of that existing research targets a different goal than *language learning*.

Where Should We Start? Concretely, many in our community are already pursuing the move from WS2 to WS3 by rethinking our existing tasks and investigating whether their semantics can be expanded and grounded. Importantly, this is not a new trend (Chen and Mooney, 2008; Feng and Lapata, 2010; Bruni et al., 2014; Lazaridou et al., 2016), and task and model design can fail to require sensory perception (Thomason et al., 2019a).

However, research and progress in multimodal, grounded language understanding has accelerated dramatically in the last few years.

Elliott et al. (2016) takes the classic problem of machine translation and reframes it in the context of visual observations (Elliott and Kádár, 2017). Wang et al. (2019b) extends this paradigm by exploring machine translation with videos serving as an interlingua, and Regneri et al. (2013) introduced a foundational dataset which aligns text descriptions and semantic annotations of actions with videos. Even core tasks like syntax can be informed with visual information. Shi et al. (2019) investigate the role of visual information in syntax acquisition. This may prove necessary to learn headedness (e.g. choosing the main verb vs the more common auxiliary as the root of a sentence) (Bisk and Hockenmaier, 2015). Relatedly, Ororbia et al. (2019) investigate the benefits of visual context for language modeling.

Collaborative games have long served as a testbed for studying language (Werner and Dyer, 1991). Recently, Suhr et al. (2019a) introduced a testbed for evaluating language understanding in the service of a shared goal, and Andreas and Klein (2016) use a simpler visual paradigm for studying pragmatics. A communicative goal can also be used to study the emergence of language. Lazaridou et al. (2018) evolves agents with language that aligns to perceptual features of the environment. These paradigms can help us examine how inductive biases and environmental pressures can build towards socialization (WS5).

Most of this research provides resources such as data, code, simulators and methodology for evaluating the multimodal content of linguistic representations (Silberer and Lapata, 2014; Bruni et al., 2012). Moving forward, we would like to encourage a broader re-examination of how NLP researchers frame the relationship between meaning and context. We believe that the time is ripe to begin a deeper dive into grounded learning, embodied learning, and ultimately social participation. With recent deep learning representations, researchers can more easily integrate additional signals into the learned meanings of language than just word tokens. We particularly advocate for the homegrown creation of such merged representations, rather than waiting for the representations and data to come to NLP from other fields.

> These problems include the need to bring meaning and reasoning into systems that perform natural language processing, the need to infer and represent causality, the need to develop computationally-tractable representations of uncertainty and the need to develop systems that formulate and pursue long-term goals.
>
> *Michael Jordan (Artificial intelligence – the revolution hasn't happened yet, 2019)*

7 Conclusions

Our proposed World Scopes are steep steps, and it is possible that WS5 is AI-complete. WS5 implies a persistent agent experiencing time and a personalized set of experiences. With few exceptions (Carlson et al., 2010), machine learning models have been confined to IID datasets that do not have the structure in time from which humans draw correlations about long-range causal dependencies.

What happens if a machine is allowed to participate consistently? This is difficult to imagine testing under current evaluation paradigms for generalization. Yet, this is the structure of generalization in human development: drawing analogies to episodic memories, gathering a system through non-independent experiments.

While it is fascinating to broadly consider such far reaching futures, our goal in this call to action is more pedestrian. As with many who have analyzed the history of Natural Language Processing, its trends (Church, 2007), its maturation toward a science (Steedman, 2008), and its major challenges (Hirschberg and Manning, 2015; McClelland et al., 2019), we hope to provide momentum for a direction many are already heading. We call for and embrace the incremental, but purposeful, contextualization of language in human experience. With all that we have learned about what words can tell us and what they keep implicit, now is the time to ask: *What tasks, representations, and inductive-biases will fill the gaps?*

Computer vision and speech recognition are mature enough for novel investigation of broader linguistic contexts (Section 3). The robotics industry is rapidly developing commodity hardware and sophisticated software that both facilitate new research and expect to incorporate language technologies (Section 4). Our call to action is to encourage the community to lean in to trends prioritizing grounding and agency (Section 5), and explicitly aim to broaden the corresponding World Scopes available to our models.

Acknowledgements

Thanks to Raymond Mooney for discussions and suggestions, Paul Smolensky for discussion and disagreements, and to Catriona Silvey for help with references.

A Invitation to Reply

What we have introduced here is not comprehensive and you may not agree with our arguments. Perhaps they do not go far enough, they miss a relevant literature, or you feel they do not represent the goals of NLP/CL. To this end, we welcome both suggestions for an updated version of this manuscript, as well as response papers on alternate foci and directions for the field.

References

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

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and visual question answering. *Visual Question Answering Challenge at CVPR 2017*.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Austin, Texas. Association for Computational Linguistics.

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.

John Langshaw Austin. 1975. *How to do things with words*, volume 88. Oxford university press.

Philip Bachman, R Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. 2019. Phyre: A new benchmark for physical reasoning.

Dare A. Baldwin, Ellen M. Markman, Brigitte Bill, Renee N. Desjardins, Jane M. Irwin, and Glynnis Tidball. 1996. Infants' reliance on a social criterion for establishing word-object relations. *Child Development*, 67(6):3135–3153.

H.B. Barlow. 1989. Unsupervised learning. *Neural Computation*, 1(3):295–311.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Rachel Barr. 2013. Memory constraints on infant learning from picture books, television, and touchscreens. *Child Development Perspectives*, 7(4):205–210.

Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.*, 59:617–645.

Yoshua Bengio, RÃl'jean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Leon Bergen, Roger Levy, and Noah Goodman. 2016. Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, 9.

Yonatan Bisk and Julia Hockenmaier. 2015. Probing the linguistic strengths and limitations of unsupervised grammar induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Yonatan Bisk, Kevin Shih, Yejin Choi, and Daniel Marcu. 2018. Learning interpretable spatial operations in a rich 3d blocks world. In *Proceedings of the Thirty-Second Conference on Artificial Intelligence (AAAI-18)*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Paul Bloom. 2002. *How children learn the meanings of words*. MIT press.

Valts Blukis, Yannick Terme, Eyvind Niklasson, Ross A. Knepper, and Yoav Artzi. 2019. Learning to map natural language instructions to physical quadcopter control using simulated flight. In *3rd Conference on Robot Learning (CoRL)*.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. In *International Conference on Learning Representations*.

Peter F Brown, Peter V deSouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145, Jeju Island, Korea. Association for Computational Linguistics.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Alexandre Campeau-Lecours, Hugo Lamontagne, Simon Latour, Philippe Fauteux, Véronique Maheu, François Boucher, Charles Deguire, and Louis-Joseph Caron L'Ecuyer. 2019. Kinova modular robot arms for service robotics applications. In *Rapid Automation: Concepts, Methodologies, Tools, and Applications*, pages 693–719. IGI Global.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka, and Tom M Mitchell. 2010. Toward an architecture for never-ending language learning. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

Malinda Carpenter, Katherine Nagell, Michael Tomasello, George Butterworth, and Chris Moore. 1998. Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the society for research in child development*, pages i–174.

Joyce Y. Chai, Rui Fang, Changsong Liu, and Lanbo She. 2017. Collaborative language grounding toward situated human-robot dialogue. *AI Magazine*, 37(4):32–45.

Joyce Y. Chai, Qiaozi Gao, Lanbo She, Shaohua Yang, Sari Saba-Sadiya, and Guangyue Xu. 2018. Language to action: Towards interactive task learning with physical agents. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*.

Soravit Changpinyo, Wei-Lun Chao, and Fei Sha. 2017. Predicting visual exemplars of unseen classes for zero-shot learning. In *ICCV*.

Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. 2016. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*, pages 52–68, Cham. Springer International Publishing.

Eugene Charniak. 1977. A framed painting: The representation of a common sense knowledge fragment. *Cognitive Science*, 1(4):355–394.

Chun-Yen Chen, Dian Yu, Weiming Wen, Yi Mang Yang, Jiaping Zhang, Mingyang Zhou, Kevin Jesse, Austin Chau, Antara Bhowmick, Shreenath Iyer, et al. 2018. Gunrock: Building a human-like social bot by leveraging large scale real user data. *Alexa Prize Proceedings*.

David L. Chen and Raymond J. Mooney. 2008. Learning to sportscast: A test of grounded language acquisition. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, Helsinki, Finland.

SF Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Association for Computational Linguistics*, pages 310–318. Association for Computational Linguistics.

Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. 2019. Babyai: First steps towards grounded language learning with a human in the loop. In *ICLR'2019*.

Noam Chomsky. 1981. *Lectures on Government and Binding*. Mouton de Gruyter.

Kenneth Church. 2007. A pendulum swung too far. *Linguistic Issues in Language Technology âĂŞ LiLT*, 2.

L. Coleman and P. Kay. 1981. The english word “lie". *Linguistics*, 57.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *ICML*.

Erwin Coumans and Yunfei Bai. 2016–2019. Pybullet, a python module for physics simulation for games, robotics and machine learning. `http://pybullet.org`.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2054–2063.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1988. Improving information retrieval with latent semantic indexing. In *Proceedings of the 51st Annual Meeting of the American Society for Information Science 25*, page 36âĂŞ40.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Gerald Dejong. 1981. Generalizations based on explanations. In *IJCAI*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4529–4538.

Susan T. Dumais. 2004. Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1):188–230.

Robin IM Dunbar. 1993. Coevolution of neocortical size, group size and language in humans. *Behavioral and brain sciences*, 16(4):681–694.

Yanai Elazar, Abhijit Mahabal, Deepak Ramachandran, Tania Bedrax-Weiss, and Dan Roth. 2019. How large are lions? inducing distributions over quantitative attributes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3973–3983.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Workshop on Vision and Langauge at ACL '16*.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 130–141.

J Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 897–906, Honolulu, Hawaii. Association for Computational Linguistics.

Susan Ervin-Tripp. 1973. Some strategies for the first two years. In Timothy E. Moore, editor, *Cognitive Development and Acquisition of Language*, pages 261 – 286. Academic Press, San Diego.

Ali Farhadi, M Hejrati, M Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *European Conference on Computer Vision*. Springer.

Yansong Feng and Mirella Lapata. 2010. Topic models for image annotation and text illustration. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 831–839, Los Angeles, California. Association for Computational Linguistics.

J. R. Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*.

Susan T Fiske and Shelley E Taylor. 1991. *Social cognition*. Mcgraw-Hill Book Company.

Cliff Fitzgerald. 2013. Developing baxter. In *2013 IEEE Conference on Technologies for Practical Robot Applications (TePRA)*.

W. Nelson Francis. 1964. A standard sample of present-day english for use with digital computers. *Report to the U.S Office of Education on Cooperative Research Project No. E-007*.

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning semantic hierarchies via word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1199–1209.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating NLP Models via Contrast Sets. *arXiv:2004.02709*.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking nli systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655.

I J Good. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40:237–264.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. 2018. Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, pages 4089–4098.

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

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

Zellig S Harris. 1954. Distributional structure. *Word*, 10:146–162.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343.

Susan J. Hespos and Elizabeth S. Spelke. 2004. Conceptual precursors to language. *Nature*, 430.

G. E. Hinton, J. L. McClelland, and D. E. Rumelhart. 1986. Distributed representations. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*.

Geoffrey E. Hinton. 1990. Preface to the special issue on connectionist symbol processing. *Artificial Intelligence*, 46(1):1 – 4.

Julia Hirschberg and Christopher D Manning. 2015. Advances in natural language processing. *Science*, 349(6245):261–266.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *ICLR 2020*.

Daniel L. James and Risto Miikkulainen. 1995. Sardnet: A self-organizing feature map for sequences. In *Advances in Neural Information Processing Systems 7 (NIPS'94)*, pages 577–584, Denver, CO. Cambridge, MA: MIT Press.

Michael I Jordan. 2019. Artificial intelligence – the revolution hasn't happened yet. *Harvard Data Science Review*.

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

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-ff of m-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.

Teuvo Kohonen. 1984. *Self-Organization and Associative Memory*. Springer.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

George Lakoff. 1973. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic*, 2:458–508.

George Lakoff. 1980. *Metaphors We Live By*. University of Chicago Press.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Ronald W Langacker. 1987. *Foundations of cognitive grammar: Theoretical prerequisites*, volume 1. Stanford university press.

Ronald W Langacker. 1991. *Foundations of Cognitive Grammar: descriptive application. Volume 2*, volume 2. Stanford university press.

Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. 2018. Emergence of linguistic communication from referential games with symbolic and pixel input. In *Internationl Conference on Learning Representations*.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language. In *ICLR 2017*.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2016. The red one!: On learning to refer to things based on discriminative properties. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 213–218, Berlin, Germany. Association for Computational Linguistics.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5871–5876, Hong Kong, China. Association for Computational Linguistics.

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

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019a. VisualBERT: A Simple and Performant Baseline for Vision and Language. In *Work in Progress*.

Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Mingyang Chen, Ze Ma, Shiyi Wang, Hao-Shu Fang, and Cewu Lu. 2019b. HAKE: Human Activity Knowledge Engine. *arXiv:1904.06539*.

Ling-Yi Lin, Rong-Ju Cherng, and Yung-Jung Chen. 2017. Effect of touch screen tablet use on fine motor development of young children. *Physical & Occupational Therapy In Pediatrics*, 37(5):457–467. PMID: 28071977.

Changsong Liu and Joyce Yue Chai. 2015. Learning to mediate perceptual differences in situated human-robot dialogue. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 2288–2294.

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

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *11 pages, 5 figures*.

Li Lucy and Jon Gauthier. 2017. Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. *arXiv preprint arXiv:1705.11168*.

Matt MacMahon, Brian Stankiewicz, and Benjamin Kuipers. 2006. Walk the talk: Connecting language, knowledge, and action in route instructions. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-2006)*, Boston, MA, USA.

Gary Marcus and Ernest Davis. 2019. *Rebooting AI: Building Artificial Intelligence We Can Trust*. Pantheon.

Mitchell P Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19:313–330.

Lara J Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark O Riedl. 2018. Event representations for automated story generation with deep neural nets. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Cynthia Matuszek. 2018. Grounded language learning: Where robotics and nlp meet (early career spotlight). In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, Stockholm, Sweden.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6297–6308.

James L. McClelland, Felix Hill, Maja Rudolph, Jason Baldridge, and Hinrich SchÃČÅŠtze. 2019. Extending Machine Language Models toward Human-Level Language Understanding. *arXiv:1912.05877*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26.

Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alexander C. Berg, Tamara L. Berg, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *European Chapter of the Association for Computational Linguistics (EACL)*.

Tom M Mitchell. 1980. *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research âĂę.

Raymond J. Mooney. 2008. Learning to connect language and perception. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI)*, pages 1598–1601, Chicago, IL. Senior Member Paper.

Raymond J Mooney and Gerald Dejong. 1985. Learning schemata for natural language processing. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence (IJCAI-85)*, pages 681 – 687.

Roozbeh Mottaghi, Mohammad Rastegari, Abhinav Gupta, and Ali Farhadi. 2016. "what happens if..." learning to predict the effect of forces in images. In *Computer Vision – ECCV 2016*, pages 269–285, Cham. Springer International Publishing.

Adithyavairavan Murali, Tao Chen, Kalyan Vasudev Alwala, Dhiraj Gandhi, Lerrel Pinto, Saurabh Gupta, and Abhinav Gupta. 2019. Pyrobot: An open-source robotics framework for research and benchmarking. *arXiv preprint arXiv:1906.08236*.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. 2018. Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2392–2400.

P Norvig and Stuart Russel. 2002. *Artificial Intelligence: A Modern Approach*. Prentice Hall.

NVIDIA. 2019. NVIDIA Isaac software development kit. https://developer.nvidia.com/isaac-sdk. Accessed 2019-12-09.

Elinor Ochs. 1993. Constructing social identity: A language socialization perspective. *Research on language and social interaction*, 26(3):287–306.

William O'Grady. 2005. *How Children Learn Language*. Cambridge University Press.

Veronica Ornaghi, Jens Brockmeier, and Ilaria Grazzani Gavazzi. 2011. The role of language games in children's understanding of mental states: A training study. *Journal of cognition and development*, 12(2):239–259.

Alexander Ororbia, Ankur Mali, Matthew Kelly, and David Reitter. 2019. Like a baby: Visually situated neural language acquisition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5127–5136, Florence, Italy. Association for Computational Linguistics.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc-Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534.

Haoruo Peng, Daniel Khashabi, and Dan Roth. 2015. Solving hard coreference problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Don Perlis. 2016. Five dimensions of reasoning in the wild. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018b. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

John R Pierce. 1969. Whither speech recognition? *The journal of the acoustical society of america*, 46(4B):1049–1051.

Nicolas Pinto, David Doukhan, James J DiCarlo, and David D Cox. 2009. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS computational biology*, 5(11):e1000579.

Jordan B. Pollack. 1987. *On Connectionist Models of Natural Language Processing*. Ph.D. thesis, University of Illinois.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

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

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you donâĂŹt know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. 2019. Stand-alone self-attention in vision models. *CoRR*, abs/1906.05909.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. 2016. Xnor-net: Imagenet classification using binary convolutional neural networks. In *ECCV*.

Joseph Redmon and Ali Farhadi. 2018. [link].

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics (TACL)*, 1:25–36.

Stephanie Rosenthal, Joydeep Biswas, and Manuela Veloso. 2010. An effective personal mobile robot agent through symbiotic human-robot interaction. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, pages 915–922. International Foundation for Autonomous Agents and Multiagent Systems.

Corby Rosset. 2020. Microsoft.com.

Jacqueline Sachs, Barbara Bard, and Marie L Johnson. 1981. Language learning with restricted input: Case studies of two hearing children of deaf parents. *Applied Psycholinguistics*, 2(1):33–54.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4462–4472, Hong Kong, China. Association for Computational Linguistics.

Rosario Scalise, Jesse Thomason, Yonatan Bisk, and Siddhartha Srinivasa. 2019. Improving robot success detection using static object data. In *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Roger C. Schank and Robert P. Abelson. 1977. *Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures*. L. Erlbaum, Hillsdale, NJ.

Marten van Schijndel, Aaron Mueller, and Tal Linzen. 2019. Quantity doesn't buy quality syntax with neural language models. *arXiv preprint arXiv:1909.00111*.

Iulian V Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, et al. 2017. A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. 2019. Visually grounded neural syntax acquisition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1842–1861, Florence, Italy. Association for Computational Linguistics.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using gpu model parallelism. *arXiv preprint arXiv:1909.08053*.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2019. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. *arXiv:1912.01734*.

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

Jivko Sinapov, Connor Schenck, and Alexander Stoytchev. 2014. Learning relational object categories using behavioral exploration and multimodal perception. In *IEEE International Conference on Robotics and Automation*.

Paul Smolensky. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46:159–216.

Richard Socher, Brody Huval, Christopher Manning, and Andrew Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Mark Steedman. 2008. Last words: On becoming a discipline. *Computational Linguistics*, 34(1):137–144.

Greg J Stephens, Lauren J Silbert, and Uri Hasson. 2010. Speaker–listener neural coupling underlies successful communication. *Proceedings of the National Academy of Sciences*, 107(32):14425–14430.

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

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019b. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. 2019a. Contrastive bidirectional transformer for temporal representation learning.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019b. VideoBERT: A Joint Model for Video and Language Representation Learning.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624. International World Wide Web Conferences Steering Committee.

Mingxing Tan and Quoc V. Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *ICML*.

Yee-Whye Teh. 2006. A hierarchical bayesian language model based on pitman-yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 985–992, Sydney, Australia.

Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. 2020. Robots that use language. *The Annual Review of Control, Robotics, and Autonomous Systems*, 15.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the National Conference on Artificial Intelligence*.

Jesse Thomason, Daniel Gordon, and Yonatan Bisk. 2019a. Shifting the baseline: Single modality performance on visual navigation & qa. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2019b. Vision-and-Dialog Navigation. In *Conference on Robotic Learning*.

Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Justin Hart, Peter Stone, and Raymond J. Mooney. 2017. Opportunistic active learning for grounding natural language descriptions. In *Proceedings of the 1st Annual Conference on Robot Learning (CoRL)*.

Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Nick Walker, Yuqian Jiang, Harel Yedidsion, Justin Hart, Peter Stone, and Raymond J. Mooney. 2020. Jointly improving parsing and perception for natural language commands through human-robot dialog. *The Journal of Artificial Intelligence Research (JAIR)*, 67.

Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond J Mooney. 2016. Learning multi-modal grounded linguistic semantics by playing "i spy". In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI-16)*, pages 3477–3483, New York City.

Emanuel Todorov, Tom Erez, and Yuval Tassa. 2012. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE.

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

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden. Association for Computational Linguistics.

Alan M Turing. 1950. Computing machinery and intelligence. *Mind*.

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Gabriella Vigliocco, Pamela Perniss, and David Vinson. 2014. Language as a multimodal phenomenon: implications for language learning, processing and evolution.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019b. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *The IEEE International Conference on Computer Vision (ICCV)*.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019c. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Ronald Wardhaugh. 2011. *An introduction to sociolinguistics*, volume 28. John Wiley & Sons.

Joseph Weizenbaum. 1966. ElizaâĂŤa computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Lloyd R Welch. 2003. Hidden markov models and the baum-welch algorithm. *IEEE Information Theory Society Newsletter*, 53(4):1–24.

Gregory M Werner and Michael G Dyer. 1991. Evolution of communication in artificial organisms. *ALife*, page 15.

Terry Winograd. 1972. Understanding natural language. *Cognitive psychology*.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Macmillan.

Ludwig Wittgenstein. 1958. *The blue and brown books*. Basil Blackwell.

Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. 2020. SAPIEN: A simulated part-based interactive environment. In *Computer Vision and Pattern Recognition (CVPR)*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Conference on Computer Vision and Pattern Recognition*.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019a. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019b. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Elizabeth Clark, Lianhui Qin, Ali Farhadi, and Yejin Choi. 2020. Evaluating machines by their real-world language use. *arXiv preprint arXiv:2004.03607*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019c. Defending against neural fake news. In *Thirty-third Conference on Neural Information Processing Systems*.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2019. Unified vision-language pre-training for image captioning and vqa. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

George Zipf. 1932. *Selected Studies of the Principle of Relative Frequency in Language*. Harvard University Press.