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

Abstract

Language understanding research is held back by a failure to relate language to the physical world it describes and to the social interactions it facilitates. Despite the incredible effectiveness of language processing models to tackle tasks after being trained on text alone, successful linguistic *communication* relies on a shared experience of the world. It is this shared experience that makes utterances meaningful.

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-only corpora requires the parallel tradition of research on the broader physical and social context of language to address the deeper questions of communication.

Improvements in hardware and data collection have galvanized progress in NLP across many benchmark tasks. Impressive performance has been achieved in language modeling (Radford et al., 2019; Zellers et al., 2019b; Keskar et al., 2019) and span-selection question answering (Devlin et al., 2019; Yang et al., 2019b; Lan et al., 2020) through massive data and massive models. With models exceeding human performance on such tasks, now is an excellent time to reflect on a key question:

Where is NLP going?

In this paper, we consider how the data and world a language learner is exposed to define and constrains the scope of that learner's semantics. Meaning does not arise from the statistical distribution of words, but from their use by people to communicate. Many of the assumptions and understandings on which communication relies lie outside of text. We must consider what is missing from models trained solely on text corpora, even when those corpora are meticulously annotated or Internet-scale.

> 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)*

You can't learn language from the radio. Nearly every NLP course will at some point make this claim. The futility of learning language from linguistic signal alone is intuitive, and mirrors the belief that humans lean deeply on non-linguistic knowledge (Chomsky, 1965, 1980). However, as a field we attempt this futility: trying to learn language from the *Internet*, which stands in as the modern radio to deliver limitless language. In this piece, we argue that 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.

Drawing inspiration from previous work in NLP, Cognitive Science, and Linguistics, we propose the notion of a World Scope (WS) as a lens through which to audit progress in NLP. We describe five WSs, and note that most trending work in NLP operates in the second (Internet-scale data).

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

- WS1. Corpus *(our past)*
- WS2. Internet *(most of current NLP)*
- WS3. Perception *(multimodal NLP)*
- WS4. Embodiment
- WS5. Social

These World Scopes go beyond text to consider the contextual foundations of language: grounding, embodiment, and social interaction. We describe a brief history and ongoing progression of how contextual information can factor into representations and tasks. We conclude with a discussion of how

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 8718–8735,
November 16–20, 2020.

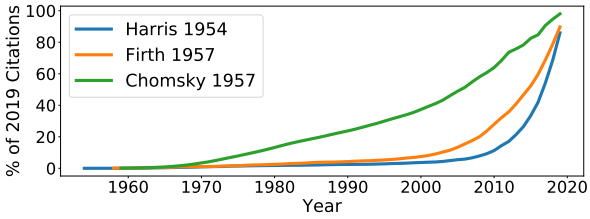
this integration can move the field forward. We believe this World Scope framing serves as a roadmap for truly contextual language understanding.

1 WS1: Corpora and Representations

The story of data-driven language research begins with the corpus. The Penn Treebank (Marcus et al., 1993) is the canonical example of a clean subset of naturally generated language, processed and annotated for the purpose of studying representations. Such corpora and the model representations built from them exemplify WS1. Community energy was initially directed at finding formal linguistic *structure*, such as recovering syntax trees. Recent success on downstream tasks has not required such explicitly annotated signal, leaning instead on unstructured fuzzy representations. These representations span from dense word vectors (Mikolov et al., 2013) to contextualized pretrained representations (Peters et al., 2018; Devlin et al., 2019).

Word representations have a long history predating the recent success of deep learning methods. Outside of NLP, philosophy (Austin, 1975) and linguistics (Lakoff, 1973; Coleman and Kay, 1981) recognized that meaning is flexible yet structured. Early experiments on neural networks trained with sequences of words (Elman, 1990; Bengio et al., 2003) suggested that vector representations could capture both syntax and semantics. Subsequent experiments with larger models, documents, and corpora have demonstrated that representations learned from text capture a great deal of information about meaning in and out of context (Collobert and Weston, 2008; Turian et al., 2010; Mikolov et al., 2013; McCann et al., 2017).

The intuition of such embedding representations, that context lends meaning, has long been acknowledged (Firth, 1957; Turney and Pantel, 2010). Earlier on, discrete, hierarchical representations, such as agglomerative clustering guided by mutual information (Brown et al., 1992), were constructed with some innate interpretability. A word's position in such a hierarchy captures semantic and syntactic distinctions. When the Baum–Welch algorithm (Welch, 2003) is applied to unsupervised Hidden Markov Models, it assigns a class distribution to every word, and that distribution is a partial representation of a word's "meaning." If the set of classes is small, syntax-like classes are induced; if the set is large, classes become more semantic. These representations are powerful in that they capture linguistic intuitions without supervision, but they are constrained by the structure they impose with respect to the number of classes chosen.



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."

The intuition that meaning requires a large context, that *"You shall know a word by the company it keeps."* – Firth (1957), manifested early via Latent Semantic Indexing/Analysis (Deerwester et al., 1988, 1990; Dumais, 2004) and later in the generative framework of Latent Dirichlet Allocation (Blei et al., 2003). LDA represents a document as a bag-of-words conditioned on latent topics, while LSI/A use singular value decomposition to project a co-occurrence matrix to a low dimensional word vector that preserves locality. These methods discard sentence structure in favor of the document.

Representing words through other words is a comfortable proposition, as it provides the illusion of definitions by implicit analogy to thesauri and related words in a dictionary definition. However, the recent trends in deep learning approaches to language modeling favor representing meaning in fixed-length vectors with no obvious interpretation. The question of *where* meaning resides in "connectionist" systems like Deep Neural Networks is an old one (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*

In connectionism, words were no longer defined over interpretable dimensions or symbols, which were perceived as having intrinsic meaning. The tension of modeling symbols and distributed representations is 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.

The Brown Corpus (Francis, 1964) and Penn Treebank (Marcus et al., 1993) defined context and structure in NLP for decades. Only relatively recently (Baroni et al., 2009) has the cost of annotations decreased enough, and have large-scale web-crawls become viable, to enable the introduction of more complex text-based tasks. This transition to larger, unstructured context (WS2) induced a richer semantics than was previously believed possible under the distributional hypothesis.

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.[1] 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. But, a large and noisy text corpus is still a text corpus.

This move towards using large scale raw data has led to substantial advances in performance on existing and novel community benchmarks (Devlin et al., 2019; Brown et al., 2020). Scale in data and modeling has demonstrated that a single representation can discover both rich syntax and semantics without our help (Tenney et al., 2019). This change is perhaps best seen in transfer learning enabled by representations in deep models. Traditionally, transfer learning relied on *our* understanding of model classes, such as English grammar. Domain adaptation simply required sufficient data to capture lexical variation, by assuming most higher-level structure would remain the same. Unsupervised representations today capture deep associations across multiple domains, and can be used successfully transfer knowledge into surprisingly diverse contexts (Brown et al., 2020).

These representations require scale in terms of both 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. Where Peters et al. (2018) introduced ELMo with $\sim 10^8$ parameters, Transformer models (Vaswani et al., 2017) have continued to scale by orders of magnitude between papers (Devlin et al., 2019; Radford et al., 2019; Zellers et al., 2019b) to $\sim 10^{11}$ (Brown et al., 2020).

Current models are the next (impressive) step in language modeling 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.

How can models yield both "impressive results" and "diminishing returns"? Language modeling—the modern workhorse of neural NLP systems—is a canonical example. Recent pretraining literature has produced results that few could have predicted, crowding leaderboards with "super-human" accuracy (Rajpurkar et al., 2018). However, there are diminishing returns. For example, on 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%). When adding another *order of magnitude* of parameters (175B) Brown et al. (2020) gain 8 percentage-points, impressive but still leaving 25% unsolved. Continuing to expand hardware, data sizes, and financial compute cost by orders of magnitude will yield further gains, but the slope of the increase is quickly decreasing.

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., 2019a). The extent to which they capture deeper notions of contextual meaning remains an open question. Past work has found that pretrained word and sentence representations fail to capture many grounded features of words (Lucy and Gauthier, 2017) and sentences, and current NLU sys-

[1] A parallel discussion would focus on the hardware required to enable advances to higher World Scopes. Playstations (Pinto et al., 2009) and then GPUs (Krizhevsky et al., 2012) made many WS2 advances possible. Perception, interaction, and robotics leverage other new hardware.

tems fail on the thick tail of experience-informed inferences, such as hard coreference problems (Peng et al., 2015). "I parked my car in the compact parking space because it looked (big/small) enough." still presents problems for text-only learners.

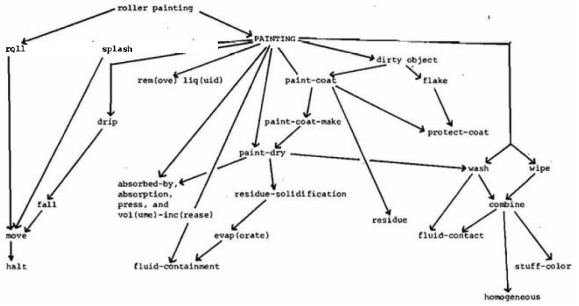
As text pretraining schemes seem to be reaching the point of diminishing returns, even for some syntactic phenomena (van Schijndel et al., 2019), we posit that other forms of supervision, such as multimodal perception (Ilharco et al., 2019), are necessary to learn the remaining aspects of meaning in context. Learning by observation should not be a purely linguistic process, since leveraging and combining the patterns of multimodal perception can combinatorially boost the amount of signal in data through cross-referencing and synthesis.

3 WS3: The World of Sights and Sounds

Language learning needs perception, because perception forms the basis for many of our semantic axioms. Learned, physical heuristics, such as the fact 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, commonly driven by mental simulation and analogy (Hofstadter and Sander, 2013). Perception is the foremost source of reporting bias. The assumption that we all see and hear the same things informs not just what we name, but what we choose to assume and leave unwritten. Further, there exists strong evidence that children require grounded sensory perception, 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, sarcasm, stress, and meaning can be implied through prosody. Further, tactile senses lend meaning, both physical (Sinapov et al., 2014; Thomason et al., 2016) and abstract, to concepts like *heavy* and *soft*. Visual perception is a rich 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



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

from an exponential number of possible how-to, text-only guides and manuals (Bisk et al., 2020) is misdirected without *some* fundamental referents to which to ground symbols. 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) and access fine-grained notions of causality, physics, and social interactions.

While the NLP community has played an important role in the history of grounding (Mooney, 2008), recently remarkable progress has taken place in the Computer Vision community. It is tempting to assume that vision models trained to identify 1,000 ImageNet classes (Russakovsky et al., 2015)[2] are limited to extracting a bag of visual words. In reality, Computer Vision has been making in-roads into complex visual, physical, and social phenomena, while providing reusable infrastructure.[3] The stability of these architectures allows for new research into more challenging world modeling. Mottaghi et al. (2016) predicts 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) models scripts and actions, and alternative unsupervised training regimes (Bachman et al., 2019) open up research towards automatic concept formation.

Advances in computer vision 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; Yagcioglu et al., 2018), natural language and visual reasoning (Suhr et al., 2019b), visual commonsense (Zellers et al., 2019a),

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

[3]Torchvision/Detectron2 include dozens of trained models.

and multilingual captioning/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 (e.g. via conceptual captions (Sharma et al., 2018)) or further broadened to include audio (Tsai et al., 2019). Vision can also help ground speech signals (Srinivasan et al., 2020; Harwath et al., 2019) to facilitate discovery of linguistic concepts (Harwath et al., 2020).

At the same time, NLP resources contributed to the success of these vision backbones. Hierarchical semantic representations emerge from ImageNet classification pretraining partially due to class hypernyms owed to that dataset's WordNet origins. 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 20K ImageNet categories (Chao et al., 2016; Changpinyo et al., 2017). Modern architectures also learn to differentiate instances within a general class, such as *face*. For example, facial recognition benchmarks require distinguishing over 10K unique faces (Liu et al., 2015). While vision is by no means "solved," benchmarks have led to off-the-shelf tools for building representations rich enough to identify tens of thousands of objects, scenes, and individuals.

A WS3 agent, having access to potentially endless hours of video data showing the intricate details of daily comings and goings, procedures, and events, reduces susceptibility to the reporting bias of WS2. An ideal WS3 agent will exhibit better long-tail generalization and understanding than any language-only system could. This generalization should manifest in existing benchmarks, but would be most prominent in a test of zero-shot circumstances, such as "Will *this* car fit through *that* tunnel?," and rarely documented behaviors as examined in script learning. Yet the WS3 agent will likely fail to answer, "Would a ceramic or paper plate make a better frisbee?" The agent has not tried to throw various objects and understand how their velocity and shape interact with the atmosphere to create lift. The agent cannot test novel hypotheses by intervention and action in the world.

> 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)*

4 WS4: Embodiment and Action

In human development, *interactive* multimodal sensory experience forms the basis of action-oriented categories (Thelen and Smith, 1996) as children learn how to manipulate their perception by manipulating their environment. Language grounding enables an agent to connect words to these action-oriented categories for communication (Smith and Gasser, 2005), but requires action to fully discover such connections. Embodiment—situated action taking—is therefore a natural next broader context.

An embodied agent, whether in a virtual world, such as a 2D Maze (MacMahon et al., 2006), a grid world (Chevalier-Boisvert et al., 2019), a simulated house (Anderson et al., 2018; Thomason et al., 2019b; Shridhar et al., 2020), or the real world (Tellex et al., 2011; Matuszek, 2018; Thomason et al., 2020; Tellex et al., 2020) must translate from language to action. 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 and weight (Thomason et al., 2017), or feedback, like muscle activations (Moro and Kennington, 2018), become available. 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 more like a banana?*

WS1 is likely not to have an answer beyond that the objects are common nouns that can both be held. WS2 may capture that oranges and baseballs both roll, but is not the deformation strength, surface texture, or relative sizes of these objects (Elazar et al., 2019). WS3 may realize 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. WS4 can appreciate the nuances of the question—the orange and baseball afford similar manipulation *because* they

have similar texture and weight, while the orange and banana both contain peels, deform, and are edible. People can reason over rich representations of common objects that these words evoke.

Planning 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 a block world concepts like "mirroring" appear (Bisk et al., 2018). Humans generalize and apply physical phenomena to abstract concepts with ease.

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 difficult.

Part of the problem is that much of the knowledge humans hold about the world is intuitive, possibly incommunicable by language, but still required to understand language. Much of this knowledge revolves around physical realities that real-world agents will encounter. Consider how many explicit and implicit metaphors are based on the idea that far-away things have little influence on manipulating local space: "a distant concern" and "we'll cross that bridge when we come to it."

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., 2020) 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).

> 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*

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

Interpersonal communication is the foundational use case of natural language (Dunbar, 1993). The physical world gives meaning to metaphors and instructions, but utterances come from a source with a purpose. Take J.L. Austin's classic example of "BULL" being written on the side of a fence in a large field (Austin, 1975). It is a fundamentally *social* inference to realize that this word indicates the presence of a dangerous creature, and that the word is written on the *opposite side* of the fence from where that creature lives.

Interpersonal dialogue as a grand test for AI is older than the term "artificial intelligence," beginning at least with Turing (1950)'s Imitation Game. Turing was careful to show how easily a naïve tester could be tricked. Framing, such as suggesting that a chatbot speaks English as a second language (Sample and Hern, 2014), can create the appearance of genuine content where there is none (Weizenbaum, 1966). This phenomenon has been noted countless times, from criticisms of Speech Recognition as "deceit and glamour" (Pierce, 1969) to complaints of humanity's "gullibility gap" (Marcus and Davis, 2019). We instead focus on why the social world is vital to *language learning*.

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). In linguistics, the usage-based theory of language acquisition suggests that constructions that are useful are the building blocks for everything else (Langacker, 1987, 1991). The *economy* of this notion of use has been the subject of much inquiry and debate (Grice, 1975). In recent years, these threads have begun to shed light on what use-cases language presents in both acquisition and its initial origins in our species (Tomasello, 2009; Barsalou, 2008), indicating the fundamental role of the social world.

WS1, WS2, WS3, and WS4 expand the factorizations of information available to linguistic meaning. allows language to be a *cause* instead of just a source of data. This is the ultimate goal for a language learner: to generate language that *does* something to the world.

Passive creation and evaluation of generated language separates generated utterances from their effects on other people, and while the latter is a rich learning signal it is inherently difficult to annotate. In order to learn the effects language has on the world, an agent must *participate* in linguistic activity, such as negotiation (Yang et al., 2019a; He et al., 2018; Lewis et al., 2017), collaboration (Chai et al., 2017), visual disambiguation (Anderson et al., 2018; Lazaridou et al., 2017; Liu and Chai, 2015), or providing emotional support (Rashkin et al., 2019). These activities require inferring mental states and social outcomes—a key area of interest in itself (Zadeh et al., 2019).

What "lame" means in terms of discriminative information is always at question: it can be defined as "undesirable," 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 "You're so lame!" 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 is key to learning that effect. Active experimentation with language starkly contrasts with the disembodied chat bots that are the focus of the current dialogue community (Roller et al., 2020; 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 whose environments are not persistent enough to learn the effects of actions.

Theory of Mind When attempting to get what we want, we confront people who have their own desires and identities. The ability to consider the feelings and knowledge of others is now commonly referred to as the "Theory of Mind" (Nematzadeh et al., 2018). This paradigm has also 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).

A series of challenges that attempt to address this fundamental aspect of communication have been introduced (Nematzadeh et al., 2018; Sap et al., 2019). These works are a great start towards deeper understanding, but static datasets can be problematic due to the risk of embedding spurious patterns and bias (de Vries et al., 2020; Le et al., 2019; Gururangan et al., 2018; Glockner et al., 2018), especially because examples where annotators cannot agree (which are usually thrown out before the dataset is released) still occur in real use cases. More flexible, dynamic evaluation (Zellers et al., 2020; Dinan et al., 2019) are a partial solution, but true persistence of identity and adaption to change are both necessary and still a long way off.

Training data in WS1-4, complex and large as it can be, 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 (Liu et al., 2016; DeVault et al., 2006). First, there is a lack of inductive bias (Martin et al., 2018). Models learn what they need to discriminate between potential labels, and it is unlikely that universal function approximators such as neural networks would ever reliably posit that people, events, and causality exist without being biased towards such solutions (Mitchell, 1980). Second, current cross entropy training losses actively discourage learning the tail of the distribution properly, as statistically infrequent events are drowned out (Pennington et al., 2014; Holtzman et al., 2020). Meanwhile, it is precisely human's ability to draw on past experience and make zero-shot decisions that AI aims to emulate.

Language in a Social Context Whenever language is used between people, it exists in a concrete social context: status, role, intention, and countless other variables intersect at a specific point (Ward-

haugh, 2011). These complexities are overlooked through selecting labels on which crowd workers agree. Current notions of ground truth in dataset construction are based on crowd consensus bereft of social context. We posit that ecologically valid 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, but initial studies have been strained by the training-validation-test set scenario and reference-backed evaluations. Collecting data about rich natural situations is often impossible. To address this gap, learning by participation, where users can freely interact with an agent, 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. Such experimentation in social intelligence is simply not possible with a fixed corpus. Once models are expected to be interacted with when tested, probing their decision boundaries for simplifications of reality and a lack of commonsense knowledge as in Gardner et al.; Kaushik et al. will become natural.

6 Self-Evaluation

We use the notion of World Scopes to make the following concrete claims:

You can't learn language ...

... from the radio (Internet). **WS2 ⊂ WS3**

> *A task learner cannot be said to be in* WS*3 if it can succeed without perception (e.g., visual, auditory).*

... from a television. **WS3 ⊂ WS4**

> *A task learner cannot be said to be in* WS*4 if the space of its world actions and consequences can be enumerated.*

... by yourself. **WS4 ⊂ WS5**

> *A task learner cannot be said to be in* WS*5 unless achieving its goals requires cooperating with a human in the loop.*

By these definitions, most of NLP research still resides in WS2. This fact 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*.

> 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)*

Where Should We Start? Many in our community are already examining phenomena in WSs 3-5. Note that research can explore higher WS phenomena without a resultant learner being in a higher WS. For example, a chatbot can investigate principles of the social world, but still lack the underlying social standing required for WS5. Next we describe four language use contexts which we believe are both research questions to be tackled and help illustrate the need to move beyond WS2.

Second language acquisition when visiting a foreign country leverages a shared, social world model that allows pointing to referent objects and miming internal states like hunger. The interlingua is physical and experiential. Such a rich internal world model should also be the goal for MT models: starting with images (Huang et al., 2020), moving through simulation, and then to the real world.

Coreference and WSD leverage a shared scene and theory of mind. To what extent are current coreference resolution issues resolved if an agent models the listener's desires and experiences explicitly rather than looking solely for adjacent lexical items? This setting is easiest to explore in embodied environments, but is not exclusive to them (e.g., TextWorld (Côté et al., 2018)).

Novel word learning from tactile knowledge and use: What is the instrument that you wear like a guitar but play like a piano? Objects can be described with both gestures and words about appearance and function. Such knowledge could begin to tackle physical metaphors that current NLP systems struggle with.

Personally charged language: How should a dialogue agent learn what is hurtful to a specific person? To someone who is sensitive about their grades because they had a period of struggle in school, the sentiment of "Don't be a fool!" can be hurtful, while for others it may seem playful. Social knowledge is requisite for realistic understanding of sentiment in situated human contexts.

Relevant recent work The move from WS2 to WS3 requires rethinking existing tasks and investigating where their semantics can be expanded and grounded. This idea is not new (Chen and Mooney, 2008; Feng and Lapata, 2010; Bruni et al., 2014; Lazaridou et al., 2016) and has accelerated in the last few years. Elliott et al. (2016) reframes machine translation with visual observations, a trend extended into videos (Wang et al., 2019b). Regneri et al. (2013) introduce a foundational dataset aligning text descriptions and semantic annotations of actions with videos. Vision can even inform core tasks like syntax (Shi et al., 2019) and language modeling (Ororbia et al., 2019). Careful design is key, as visually augmented tasks can fail to require sensory perception (Thomason et al., 2019a).

Language-guided, embodied agents invoke many of the challenges of WS4. Language-based navigation (Anderson et al., 2018) and task completion (Shridhar et al., 2020) in simulation environments ground language to actions, but even complex simulation action spaces can be discretized and enumerated. By contrast, language-guided robots that perform task completion (Tellex et al., 2014) and learning (She et al., 2014) in the real world face challenging, continuous perception and control (Tellex et al., 2020). Consequently, research in this space effectively restricts understanding to small grammars (Paul et al., 2018; Walter et al., 2013) or controlled dialog responses (Thomason et al., 2020). These efforts to translate language instructions to actions build towards using language for end-to-end, continuous control (WS4).

Collaborative games have long served as a testbed for studying language (Werner and Dyer, 1991) and emergent communication (Schlangen, 2019a; Lazaridou et al., 2018; Chaabouni et al., 2020). Suhr et al. (2019a) introduced an environment for evaluating language understanding in the service of a shared goal, and Andreas and Klein (2016) use a visual paradigm for studying pragmatics. Such efforts help us examine how inductive biases and environmental pressures build towards socialization (WS5), even if full social context is still too difficult and expensive to be practical.

Most of this research provides resources such as data, code, simulators and methodology for evaluating the multimodal content of linguistic representations (Schlangen, 2019b; Silberer and Lapata, 2014; Bruni et al., 2012). Moving forward, we encourage a broad re-examination of how NLP frames the relationship between meaning and context (Bender and Koller, 2020) and how pretraining obfuscates our ability to measure generalization (Linzen, 2020).

7 Conclusions

Our World Scopes are steep steps. WS5 implies a persistent agent experiencing time and a personalized set of experiences. confined to IID datasets that lack 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 test under current evaluation paradigms for generalization. Yet, this is the structure of generalization in human development: drawing analogies to episodic memories and gathering new data through non-independent experiments.

As with many who have analyzed the history of NLP, 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 investigation of broader linguistic contexts (WS3). The robotics industry is rapidly developing commodity hardware and sophisticated software that both facilitate new research and expect to incorporate language technologies (WS4). Simulators and videogames provide potential environments for social language learners (WS5). Our call to action is to encourage the community to lean in to trends prioritizing grounding and agency, and explicitly aim to broaden the corresponding World Scopes available to our models.

Acknowledgements

Thanks to Raymond Mooney for suggestions, Paul Smolensky for disagreements, Catriona Silvey for developmental psychology help, and to a superset of: Emily Bender, Ryan Cotterel, Jesse Dunietz, Edward Grefenstette, Dirk Hovy, Casey Kennington, Ajay Divakaran, David Schlangend, Diyi Yang, and Semih Yagcioglu for pointers and suggestions.

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.

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*. Oxford university press.

Philip Bachman, R Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems 32*.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. 2019. Phyre: A new benchmark for physical reasoning. In *Advances in Neural Information Processing Systems 32 (NIPS 2019)*.

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.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Association for Computational Linguistics (ACL)*.

Yoshua Bengio, Réjean 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, 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)*.

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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *preprint*.

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.

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.

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. 2020. Compositionality and generalization in emergent languages. In *Association for Computational Linguistics (ACL)*.

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.

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. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Noam Chomsky. 1980. *Language and learning: the debate between Jean Piaget and Noam Chomsky*. Harvard University Press.

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

Kenneth Church. 2007. A pendulum swung too far. *Linguistic Issues in Language Technology – 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*.

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

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*, pages 36 – 40.

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 *Proceedings of the 7th international joint conference on Artificial intelligence (IJCAI)*.

David DeVault, Iris Oved, and Matthew Stone. 2006. Societal grounding is essential to meaningful language use. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 747.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

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*.

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.

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.

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

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.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

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.

David Harwath, Wei-Ning Hsu, and James Glass. 2020. Learning hierarchical discrete linguistic units from visually-grounded speech. In *ICLR 2020*.

David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. 2019. Jointly discovering visual objects and spoken words from raw sensory input. *International Journal of Computer Vision*.

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.

Douglas Hofstadter and Emmanuel Sander. 2013. *Surfaces and essences: Analogy as the fuel and fire of thinking*. Basic Books.

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

Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander Hauptmann. 2020. Unsupervised multimodal neural machine translation with pseudo visual pivoting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8226–8237, Online.

Gabriel Ilharco, Yuan Zhang, and Jason Baldridge. 2019. Large-scale representation learning from visually grounded untranscribed speech. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 55–65, Hong Kong, China.

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-off for 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, Michael S. Bernstein, and Fei-Fei Li. 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. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

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. 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.

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.

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.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Association for Computational Linguistics (ACL)*.

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 *Advances in Neural Information Processing Systems*, pages 13–23.

Li Lucy and Jon Gauthier. 2017. Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 76–85, Vancouver, Canada. Association for Computational Linguistics.

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 Schütze. 2019. Extending Machine Language Models toward Human-Level Language Understanding. *arXiv:1912.05877*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg 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.

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)*.

Daniele Moro and Casey Kennington. 2018. Multimodal visual and simulated muscle activations for grounded semantics of hand-related descriptions. In *Workshop on the Semantics and Pragmatics of Dialogue*. SEMDIAL.

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.

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.

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.

Rohan Paul, Jacob Arkin, Derya Aksaray, Nicholas Roy, and Thomas M Howard. 2018. Efficient grounding of abstract spatial concepts for natural language interaction with robot platforms. *The International Journal of Robotics Research*, 37(10):1269–1299.

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.

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. 2018. Deep contextualized word representations. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

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.

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

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't 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.

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.

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.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot. In *arXiv*.

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. Turing-NLG: A 17-billion-parameter language model by Microsoft.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge.

International Journal of Computer Vision (IJCV), 115(3):211–252.

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.

Ian Sample and Alex Hern. 2014. Scientists dispute whether computer 'eugene goostman' passed turing test. *The Guardian*, 9.

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.

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*.

David Schlangen. 2019a. Grounded agreement games: Emphasizing conversational grounding in visual dialogue settings. *arXiv*.

David Schlangen. 2019b. Language tasks and language games: On methodology in current natural language processing research. *arXiv*.

Iulian V. Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, Sai Rajeshwar, Alexandre de Brebisson, Jose M. R. Sotelo, Dendi Suhubdy, Vincent Michalski, Alexandre Nguyen, Joelle Pineau, and Yoshua Bengio. 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.

Lanbo She, Shaohua Yang, Yu Cheng, Yunyi Jia, Joyce Y. Chai, and Ning Xi. 2014. Back to the blocks world: Learning new actions through situated human-robot dialogue. In *Proceedings of 15th SIGDIAL Meeting on Discourse and Dialogue*.

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.

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. 2020. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. *Computer Vision and Pattern Recognition (CVPR)*.

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.

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*.

Linda Smith and Michael Gasser. 2005. The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1-2):13–29.

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)*.

T. Srinivasan, R. Sanabria, and F. Metze. 2020. Looking enhances listening: Recovering missing speech using images. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6304–6308.

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.

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.

Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. 2019a. Contrastive bidirectional transformer for temporal representation learning. *arxiv:1906.05743*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019b. VideoBERT: A Joint Model for Video and Language Representation Learning. In *International Conference on Computer vision*.

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, Ross Knepper, Adrian Li, Daniela Rus, and Nicholas Roy. 2014. Asking for help using inverse semantics. In *Proceedings of Robotics: Science and Systems (RSS)*, Berkeley, California.

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*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy.

Esther Thelen and Linda B. Smith. 1996. *A Dynamic Systems Approach to the Development of Cognition and Action*. MIT Press.

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 Robot Learning (CoRL)*.

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 *International Joint Conference on Artificial Intelligence (IJCAI)*.

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.

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.

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.

Harm de Vries, Dzmitry Bahdanau, and Christopher Manning. 2020. Towards ecologically valid research on language user interfaces. In *arXiv*.

Matthew Walter, Sachithra Hemachandra, Bianca Homberg, Stefanie Tellex, and Seth Teller. 2013. Learning semantic maps from natural language descriptions. In *Proceedings of Robotics: Science and Systems (RSS)*, Berlin, Germany.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019a. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

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)*.

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

Joseph Weizenbaum. 1966. Eliza – a 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*.

Terry Winograd. 1971. Procedures as a representation for data in a computer program for understanding natural language. Technical report, Massachusetts Institute of Technology, Project MAC.

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)*.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1368, Brussels, Belgium.

Diyi Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. 2019a. Let's make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. 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)*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32 (NIPS 2019)*.

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, 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. 2019b. 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*.