COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 44 (2020) e12925
© 2020 Cognitive Science Society, Inc. All rights reserved.
ISSN: 1551-6709 online
DOI: 10.1111/cogs.12925

Analyzing Machine-Learned Representations: A Natural Language Case Study

Ishita Dasgupta,[a] Demi Guo,[b] Samuel J. Gershman,[c] Noah D. Goodman[d]

[a]*Departments of Psychology and Computer Science, Princeton University*
[b]*Department of Computer Science, Harvard University*
[c]*Department of Psychology and Center for Brain Science, Harvard University*
[d]*Departments of Psychology and Computer Science, Stanford University*

Received 21 August 2019; received in revised form 19 November 2020; accepted 19 November 2020

Abstract

As modern deep networks become more complex, and get closer to human-like capabilities in certain domains, the question arises as to how the representations and decision rules they learn compare to the ones in humans. In this work, we study representations of sentences in one such artificial system for natural language processing. We first present a diagnostic test dataset to examine the degree of abstract composable structure represented. Analyzing performance on these diagnostic tests indicates a lack of systematicity in representations and decision rules, and reveals a set of heuristic strategies. We then investigate the effect of training distribution on learning these heuristic strategies, and we study changes in these representations with various augmentations to the training set. Our results reveal parallels to the analogous representations in people. We find that these systems can learn abstract rules and generalize them to new contexts under certain circumstances—similar to human zero-shot reasoning. However, we also note some shortcomings in this generalization behavior—similar to human judgment errors like belief bias. Studying these parallels suggests new ways to understand psychological phenomena in humans as well as informs best strategies for building artificial intelligence with human-like language understanding.

Keywords: Representation learning; Natural language inference; Compositionality; Heuristic; Strategies; Sentence embeddings; Generalization; Test datasets

1. Introduction

Recent years have seen a vast improvement in the capabilities of artificial intelligence systems, driven primarily by developments in deep neural networks (for a review, see LeCun, Bengio, & Hinton, 2015). These have allowed artificial systems to reach human-

Correspondence should be sent to Ishita Dasgupta, Departments of Psychology and Computer Science, Princeton University, 35 Olden Street, Princeton, NJ 08540. E-mail: ishitad@princeton.edu

level performance at video games (Mnih et al., 2015), object recognition (Russakovsky et al., 2015), and voice generation (Oord et al., 2016), as well as produced impressive performance in several other domains. However, some serious concerns haunt deep learning approaches and their promise as a general solution to artificial intelligence. Many of these concerns surround the lack of structure in the representations and decision criteria these systems learn (Lake, Ullman, Tenenbaum, & Gershman, 2017; Marcus, 2018). This problem has been implicated in deep learning's data inefficiency and inability to learn abstract structure from few examples, its difficulty in utilizing hierarchical structure and thereby foster transfer between tasks and domains, as well as the challenge of integrating established prior information into deep learning systems. It also presents serious concerns about the interpretability of its representations and decision criteria, making them less dependable and risky for deployment in sensitive or highly variable domains.

All of this points to a crucial problem: How can we better understand the representations learned by these systems? Existing studies (e.g., Karpathy, Johnson, & Fei-Fei, 2015; Li, Chen, Hovy, & Jurafsky, 2015; Yosinski, Clune, Nguyen, Fuchs, & Lipson, 2015; Zeiler & Fergus, 2014) primarily use approaches inspired by neuroscience methods developed to understand the brain, for example, the statistical analysis of unit activations, and ablation studies where specific units are disconnected or deactivated. These methods promise interesting bottom-up insights into the inner workings of these systems. Cognitive science provides another set of tools to approach this problem from the top down (Ettinger, Elgohary, Phillips, & Resnik, 2018; Kádár, Chrupała, & Alishahi, 2017; McCoy, Pavlick, & Linzen, 2019; Ritter, Barrett, Santoro, & Botvinick, 2017), by decomposing cognitive processes into their computational components, building models that incorporate these components, and testing these by making predictions about behavior on carefully selected test problems that distinguish different hypotheses.

The cognitive science approach has yielded huge benefits in understanding higher level cognition in humans, a prime example of which is the human ability to learn, understand, and produce language (Chomsky & Lightfoot, 2002; Linzen, 2019). This domain exemplifies a hallmark of human intelligence: the ability, in the words of von Humboldt, to "make infinite use of finite means." Specifically, human cognitive abilities have been characterized as *systematic* (Fodor & Pylyshyn, 1988; Lake, Linzen, & Baroni, 2019)—this indicates an algebraic capacity to produce new combinations from known components. For example, when a person learns a word in a specific context as part of a particular sentence, they can immediately use this new word in an infinity of other sentences in which this word has never previously been encountered. Systematicity therefore allows humans an impressive capacity to *generalize*, transferring knowledge from one context to others. This ability requires the representations underlying this newly learned word, for example, to be abstract (not tied to specific contexts) and compositional (possible to combine with other words and sentences). The absence of systematicity in neural networks has been a recurring (and controversial) theme in cognitive science (Fodor & Pylyshyn, 1988; Lake et al., 2017). While several previous approaches have demonstrated the lack of compositionality in neural networks (Belinkov & Glass, 2019; Gershman &

Tenenbaum, 2015; Lake & Baroni, 2018), they have not focused on analyzing the representations that are in fact learned, how they can be altered, and their various properties.

In this paper, we carry out an analysis of a machine learning model for a difficult natural language processing task. In particular, we study its behavior on controlled test data to shed light on the sentence representations it learns. We discover that there is no evidence that the model uses systematic representations; instead, we find evidence of various heuristic strategies. We then investigate how these heuristics might arise. Analyses of the training distribution reveal that it is very biased, containing many unintended structural regularities that can be exploited by these much simpler heuristics. These simple rules are therefore easily acquired by the neural network, since they explain a substantial amount of variance without having to invoke a more complex systematic representation. We then carry out various augmentations to the training set and find that the system can learn some form of abstract composable representation, given the right training distribution. Finally, we investigate how these composable representations generalize. Several findings in cognitive science indicate that even humans do not always generalize entirely systematically (Evans, 2003; Evans & Perry, 1995). We find parallels between our findings and studies of human representations in terms of how systematic they are under certain circumstances, as well as in terms of when and where this systematicity breaks down. We discuss how such parallels can be useful to both cognitive science and machine learning.

We also note a caveat to using the behavior of a full model to shed light on its representations. The behavior of the model is supported by both the sentence representations learned as well as the decision function that takes these representations as input to produce task-specific output. This partially confounds the contributions of the representations and the decision function. In this paper, we keep the classifier relatively simple (detailed in Section 2.3), so the bulk of the behavior observed is driven by the representations. However, we cannot conclusively demonstrate the presence or absence of systematicity in the representation from this behavior. For example, it is possible that the representations are indeed systematic, but that the decision function fails to use this information correctly for the task at hand. Since the decision function is trained end-to-end with the representations, this is unlikely—the representations are unlikely to represent information that they cannot use for the task. Similarly, it is possible that the representations are not in fact systematic, and any systematic behavior should be attributed to the decision function. This, too, is unlikely given the relatively low expressivity of the decision function compared to that of the sentence encoder. Nonetheless, we note that from our methodology, we can only conclude whether or not the entire model displays systematicity, and this gives indirect evidence for whether or not the representations themselves are systematic.

2. Background

In this section, we review some background on the kinds of representations we will be studying (vector space embeddings of sentences). We also review the three key factors in how such embeddings are generated: the task that they are optimized for, the architecture of the model used to perform that task, and the training distribution on which performance is optimized.[1] We also discuss relevant related work on studying such representations.

2.1. Vector space embeddings

Vector space models represent items as vectors in some metric space. These have a long history in cognitive science as models of semantic representations (Beals, Krantz, & Tversky, 1968; Pereira, Gershman, Ritter, & Botvinick, 2016; Steyvers, 2006). In particular, in the domain of language, vector space models of words (also known as word embeddings) that are learned using distributional information (statistics of text corpora) have been shown to encode syntactic as well as semantic structure, and they have been used in psychological models for syntactic category acquisition (Redington, Crater, & Finch, 1998), inductive vocabulary learning (Landauer & Dumais, 1997), analogical reasoning (Rumelhart & Abrahamson, 1973), categorization (Jones & Mewhort, 2007), and high-level associative judgments (Bhatia, 2017). Modern machine learning has allowed the mining of very large datasets to produce vector space embeddings that are now commonly used as the word representations in artificial intelligence systems for natural language processing (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Pennington, Socher, & Manning, 2014).

Understanding language requires understanding not only words, but also their relations within a sentence. These relations are abstract and composable, allowing language to be combinatorially productive—with a finite set of words, one can systematically produce an infinite set of sentences simply by creating new and longer combinations of these known words. The number of sentences in a language therefore far exceeds the number of words. For this reason, generating similar vector embeddings for sentences has proven challenging. Recent papers have developed several supervised as well as unsupervised approaches to learning vector space representations of sentences using recurrent neural networks (RNNs) that are able to represent the order of words in a sentence (Conneau, Kiela, Schwenk, Barrault, & Bordes, 2017; Hill, Cho, & Korhonen, 2016; Kiros et al., 2015). These are intended to capture sentence-level semantic content, and they have been shown to perform reasonably well on transfer tasks (sentence-level semantic tasks on which the embeddings were not specifically trained). In particular, the performance of these sentence models exceeds the performance of representations that treat sentences as bags of words (BOW models)—these patently lack any order information about the words, therefore ignoring the abstract and composable relational structure at the sentence level. However, exactly what relational information between words is actually represented in such RNN sentence models is unclear. In this work, we start to shed light on this question.

2.2. Natural language inference

The sentence embeddings we analyze are trained on the natural language inference (NLI) task. The goal is to classify pairs of sentences (a premise and a hypothesis) into "entailment," "contradiction," or "neutral," depending on the semantic relation between the two sentences. This is a popular domain for studying artificial representations since it has a lot of relatively interpretable underlying structure (Ettinger et al., 2018; Glockner, Shwartz, & Goldberg, 2018; McCoy et al., 2019; Nie, Wang, & Bansal, 2019). For example, it is a simple domain in which abstract and composable relational structure is required—word-level information is not generally sufficient to perform well on this task. The premise sentence "Anne is more cheerful than Bob" contradicts the hypothesis sentence "Anne is less cheerful than Bob," but entails the hypothesis sentence "Bob is less cheerful than Anne." Here, both the hypothesis sentences have the exact same words, and they would be indistinguishable if we were just comparing the words in them. More generally, X is more Y than Z entails that Z is less Y than X, for any X, Y, and Z. In this case, the specific words used almost do not even matter, and the bulk of the information is in the relations between the words in the sentence. Encoding abstract rules like this allows us to systematically carry out NLI on combinatorially many different sentences, with different Xs, Ys, and Zs.

The human ability to carry out abstract reasoning of this sort is a richly studied topic. Some of these abilities however are so obvious that they are often simply taken for granted without formal study. For example, it is reasonable to assume that any adult human (in the absence of time pressure or cognitive load) can fairly easily process that if X is more Y than Z, then in general Z is less Y than X irrespective of the specific meanings of X, Y, and Z. In this paper, we investigate the extent to which certain machine-learned sentence embeddings can represent and use such abstract rules in NLI.

Despite the generally acknowledged power of human abstract reasoning, a number of studies indicate that humans are not perfect: Semantic content (e.g., the specific meanings of the X, Y, and Zs above) has been shown to interfere with systematic inferences in an effect often termed "belief bias" (Braine, 1978; Johnson-Laird & Steedman, 1978). This effect is especially noticeable in children (Evans & Perry, 1995), as well as adults under time pressure or cognitive load (Evans, 2013). In the last part of this paper, we discuss similarities between humans and machines in how they fail certain tests of systematicity.

2.3. Models for sentence embeddings

The sentence embeddings we study in this paper are from a highly successful NLI system, InferSent (Conneau et al., 2017). Each premise and hypothesis sentence is input to a sentence encoder as a sequence of pretrained 300-dimensional GloVe word embeddings (Pennington et al., 2014). These word embeddings already contain a lot of information about the semantic and syntactic roles of the words (for details, see Section 2.1), and therefore a large part of the lexical information is already represented. Therefore, the bulk of the work InferSent has to do is to learn and represent how these words relate to one another in a sentence to provide meanings. The sentence encoder takes in this variable length input and, after passing it through various recurrent and convolutional (for details, see Conneau et al., 2017), provides a 4,096-dimensional vector as output. This output vector serves as a sentence embedding. The same sentence encoding process is applied to both the premise and the hypothesis sentences. To make the final inference, both sentence embeddings are fed to the classifier described in Fig. 1 (with two linear fully connected 512-dimensional hidden layers) that labels each pair as entailment, neutral, or contradiction. The network is trained end-to-end with supervised learning, using a large labeled dataset for NLI (see next section for details on this dataset).

Using the behavior of a downstream classifier to gain insights into the properties of a representation is a common approach to analyzing high dimensional representations, for example in neuroscience (Hung, Kreiman, Poggio, & DiCarlo, 2005). But this approach runs the risk that some of the findings might reflect properties of the classifier rather than of the embeddings. We partially address this concern by keeping the classifier simple (using only linear layers), such that the bulk of the information about sentences and their meaning is stored in the embedding. This assumption is further justified by findings that the learned embeddings perform well on other sentence-level tasks (such as sentiment analysis, semantic textual similarity, and other NLI datasets; details in Conneau et al., 2017) by reusing the sentence encoder and retraining only the classifier that acts as the decision function for each specific task at hand. This indicates that the system does capture semantic content, and that it is primarily stored in the sentence embedding (not in the classifier), and is in a form that is easily decoded by simple downstream decision functions.

For our tasks, we replicate the procedure in Conneau et al. (2017) to obtain sentence embeddings. These are henceforth referred to as the InferSent sentence embeddings. Our trained InferSent model gives us 84.73% accuracy on validation and 84.84% accuracy on the test dataset, which is comparable to the performance of the classifier reported in Conneau et al. (2017). For comparison, we consider a BOW baseline model that averages the pretrained GloVe word embeddings for all the words in the sentence to form a sentence embedding. These BOW embeddings cannot represent abstract relational structure, since the architecture of the model used to generate them (a simple average of the word embeddings) cannot express word order. We then train a perceptron classifier (with the

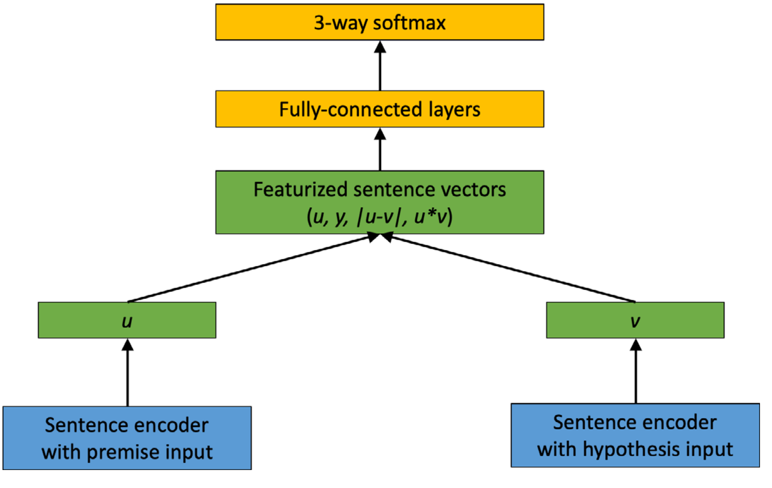

Fig. 1. InferSent architecture, figure adapted from Conneau et al. (2017).

same architecture as used in InferSent) on these embeddings to perform NLI. The two models therefore only differ in the sentence embeddings used. The BOW model achieves 53.99% accuracy on the SNLI test set (comparable to the BOW performance reported in Conneau et al., 2017).

Neural networks can act as universal function approximators (Hornik, 1991; Siegelmann & Sontag, 1995), and given sufficient capacity, they can represent any arbitrarily complex set of relations between the words in the sentence. The InferSent model has a very large capacity due to a large number of layers and hidden units (see Conneau et al., 2017), so a lot of abstract compositional structure is in theory within the representational capacity of these sentence embeddings. In this paper, we analyze how much systematic structure is actually learned and utilized for the NLI task at hand.

2.4. Training datasets

To understand sentence embeddings like the ones learned by InferSent, it is imperative to not only consider the model specifications for the system that produces them (in this case the specific end-to-end architecture of the network in InferSent), but also the learning signals it receives from the training set. For many deep learning-based methods, very little information about the structure of the task is baked into the architecture of the models—the only structure about language that it is endowed with before training are the biases that come with using an RNN as the architecture. This specifies that sentences have variable-length, sequential structure. These embedding models are therefore fairly "tabula rasa," and most of what they represent about the structure of the task (in this case NLI) is learned from training data. As elaborated in the previous section, some abstract compositional structure is within the representational capacity of the InferSent sentence embeddings—but whether or not the right structure is actually learned and represented depends largely on the training data. The significance of the training set on the representations learned by flexible deep learning methods is often not adequately considered. One contribution of this work is to highlight and analyze this issue.

InferSent was trained on the Stanford Natural Language Inference (SNLI) dataset (Bowman, Angeli, Potts, & Manning, 2015), a popular labeled dataset for NLI. SNLI consists of 550k premise–hypothesis sentence pairs, and it is balanced (consists of equal number of pairs with entailment, contradiction, and neutral relationships). The dataset was generated with a crowd-sourcing framework. Workers were presented with a scene description from a corpus of image captions that act as the premise, and asked to supply hypothesis sentences that have each of the three possible NLI relations (entailment, neutral, and contradiction) to the given premise. The freedom to produce entirely novel hypotheses leads to a rich set of sentences; however, it also leads to some artifacts that can strongly bias the representations learned by a "tabula rasa" system. We discuss these in later sections.

2.5. Related work

There has been previous work in cognitive science, studying the systematicity of neural network representations in the domain of natural language (Gershman & Tenenbaum, 2015; Lake & Baroni, 2018). These analyses, however, are often carried out on toy systems, and while they demonstrate the lack of systematicity, they do not investigate what the systems do learn or analyze the resulting representations.

There has also been interest in the field of natural language processing, to build toward a better understanding of machine-learned representations (Belinkov & Glass, 2019). Other studies concurrent with ours (Ettinger et al., 2018; Glockner et al., 2018) as well as building on our work (McCoy et al., 2019) have investigated sentence representations by analyzing behavior on controlled test sets that expose simpler word-level heuristics in NLI models. Gururangan et al. (2018), Poliak, Naradowsky, Haldar, Rudinger, and Van Durme (2018), and Tsuchiya (2018) discover and analyze incidental artifacts in SNLI that permit the success of heuristic strategies. Previous work (Bentivogli et al., 2016; Lai & Hockenmaier, 2014) has also studied the statistical structure of other datasets for NLI. Kang, Khot, Sabharwal, and Hovy (2018) and Jia and Liang (2017), among others, study adversarially augmented training in a natural language setting. None of these approaches, however, manipulate the properties of the resulting learned representations, or bridge these insights with our understanding of systematic generalization in human cognition.

3. A test dataset of minimal cases: The Comparisons dataset

Our goal is to understand the representations and decision criteria learned by InferSent, in particular how much systematic relational information they encode and utilize—do they represent abstract rules for the ways words combine to give meaning to sentences? In the machine learning literature on natural language processing, any performance above the BOW baseline (that only receives the words in the sentence with no order information) is often seen as proof of the encoding and utilization of relational information. However, this is an unwarranted conclusion—the BOW baseline usually receives only averaged word vectors for the sentence, and therefore also loses some of the lexical information. It often does not actually reach the best possible performance with only the words. Performance above this baseline therefore does not license the conclusion that relational information is being encoded and used at all. A central goal of this paper is to better test whether sentence representations encode abstract, systematic rules about the relations between words in a sentence.

Here, we pursue an alternative approach, inspired by traditions in cognitive psychology and psycholinguistics of building diagnostic test sets to investigate the underlying representations and decision rules. The goal is to generate a set of sentence pairs such that encoding the relations between words (in addition to the words themselves) is *required* to correctly classify them into the three NLI classes. Diagnostic test datasets such as these, that posit a hard baseline for performance without relational information, provide a more foolproof way to test whether such information is being used.

We considered pairs of sentences such that the NLI relation between the sentences can be changed without changing any of the words in the sentence, only their order. We generated our test dataset using comparisons as these are easy to fit into the NLI framework, and they yield many simple examples of sentence pairs that require more than word-level data to understand. For example, the premise sentence "The woman is more cheerful than the man" contradicts one hypothesis sentence, "The woman is less cheerful than the man," but entails another hypothesis sentence, "The man is less cheerful than the woman." Since both hypothesis sentences have the exact same words, they would be indistinguishable if we were just comparing their BOW representations. Therefore, a model based only on the words, and not considering the relations between them, would at most get one of the two classifications right. This caps the BOW performance at 50% , and some relational rules must be learned to perform above this baseline.

Generation of several such sentence pairs can be easily automated. We considered three subtypes, described below and summarized in Table 1. The entire dataset consisted of 14,670 sentence pairs of each kind, giving a total size of 44,010 sentence pairs.

3.1. *Same type*

Premise–Hypothesis pairs differ only in the order of the words.
Premise: The woman is more cheerful than the man.
Hypothesis: The man is more cheerful than the woman.
CONTRADICTION
Premise: The woman is more cheerful than the man.
Hypothesis: The woman is more cheerful than the man.
ENTAILMENT

3.2. *More-less type*

Premise–Hypothesis pairs differ by whether they contain the words "more" or "less."
Premise: The woman is more cheerful than the man.
Hypothesis: The woman is less cheerful than the man.
CONTRADICTION
Premise: The woman is more cheerful than the man.
Hypothesis: The man is less cheerful than the woman.
ENTAILMENT

3.3. *Not type*

Premise–Hypothesis pairs differ by whether they contain the word "not."
Premise: The woman is more cheerful than the man.

Table 1
Rules in Comparisons dataset for premise: X is more Y than Z

Type	Entailment Hypothesis	Contradiction Hypothesis
Same	X is more Y than Z	Z is more Y than X
More-Less	Z is less Y than X	X is less Y than Z
Not	Z is not more Y than X	X is not more Y than Z

Hypothesis: The woman is not more cheerful than the man.
CONTRADICTION
Premise: The woman is more cheerful than the man.
Hypothesis: The man is not more cheerful than the woman.
ENTAILMENT

To facilitate comparison with the SNLI dataset, we ensured that the vocabulary distribution of our Comparisons dataset is similar to the original SNLI training dataset.[2] This ensured that we are only manipulating the relational structure of the test set, and poor performance cannot be attributed to not having experienced the specific words before.

4. Testing the sentence embeddings

We tested the two classifiers based on two different sentence embeddings (the InferSent sentence embeddings and the BOW sentence embeddings) on the constructed test set (the Comparisons dataset, Table 1). Both of these classifiers were trained for the same task (NLI), on the same training dataset (SNLI), with the same classifier architecture, and differed only in the model used to generate the underlying sentence representations. The InferSent embeddings had access to word order, while the BOW embeddings did not (for details, see Section 2.3). The overall performance of each of the two classifiers on the Comparisons dataset are given in Table 2, and they are analyzed in greater detail in the following sections.

4.1. Performance of bag of words

We found that the BOW embeddings make classifications that are exactly symmetric across the two true labels (entailment and contradiction) in each task (rows in Fig. 2). This is expected since the sentence pairs with one label are just permuted versions of the sentence pairs with the other label. Therefore, BOW cannot distinguish them and necessarily classifies both of them the same way. This also ensures that the performance is capped at 50%. Asymmetry between the classifications of the two categories can occur only when relational information is encoded in the sentence embedding.

Considering the aggregate performance of BOW in Table 2, we found that performance, particularly on the "more/less" type subset of the test dataset (30.24%), was significantly below 50%. This highlights the trouble with using BOW embeddings as a

Table 2
Performance on the Comparisons dataset

Type	BOW	InferSent
Same	50.0	50.37
More/less	30.24	50.35
Not	48.98	45.24

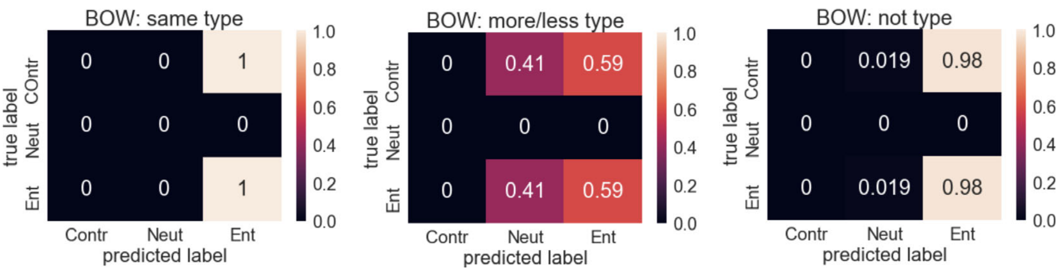

Fig. 2. BOW embedding confusion matrices, with normalized rows.

baseline for the encoding and use of relational information. Up to 50% performance is achievable on this dataset without using any relational information; therefore performance above the BOW baseline of 30.24% does not necessarily imply the use of relational information.

4.2. Performance of InferSent

The performance of the InferSent embeddings was slightly asymmetric (Fig. 3), indicating that it was able to distinguish sentences slightly, based on relational information. Yet overall the InferSent embeddings were extremely poor at this task (Table 2), achieving performances slightly above 50% for two of the three subtypes of sentence pairs in the Comparisons dataset, and even less than 50% in a third subtype. This indicates that InferSent embeddings do not correctly encode and utilize the kinds of abstract relational rules we tested with the Comparisons dataset.

However, InferSent's performance on another test dataset (the SNLI test dataset) is as high as 84%—so it is clearly encoding some relevant information about NLI. Further, a quick glance at Fig. 3 indicates that InferSent does not respond randomly to the queries in our Comparisons dataset, but rather in some structured (though incorrect) way. Rather than simply conclude that InferSent embeddings are not systematic and leaving things at that, we can study patterns in the incorrect classifications made to better understand the underlying representations and decision rules. Since our test dataset is highly structured, it allows a controlled way to generate and test hypotheses about the heuristic representations and decision rules InferSent implements.

Aside from isolating and characterizing these heuristics, it is also instructive to consider how InferSent might come to encode them in the first place. To answer this, we look to the study of heuristic strategies in humans. The theory of ecological rationality (Simon, 1991; Todd & Gigerenzer, 2007) posits that a system can exploit structural regularities in its learning environment by using heuristics that achieve close to optimal performance in that specific environment. While there might be several predictive cues that permit good performance in a given environment, ecological rationality suggests that intelligent systems will pick up on the cues that allow the "simplest" heuristics, heuristics that can be much simpler than the most general strategy that performs well in all environments. Heuristics that leverage these (incidental) structural regularities are therefore

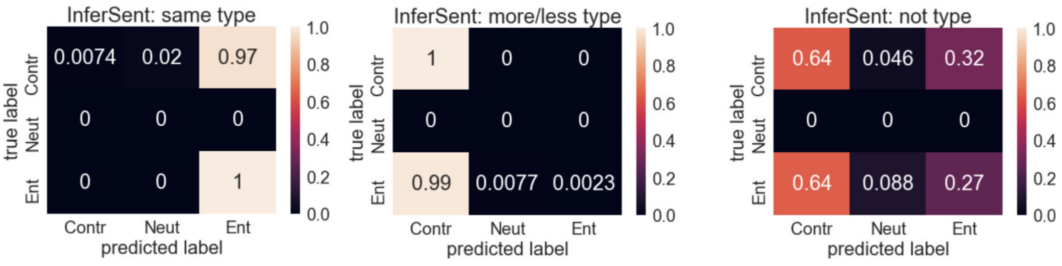

Fig. 3. InferSent embedding confusion matrices, with normalized rows.

termed "ecologically valid" in that environment. This suggests that we can better understand how heuristic strategies might arise in InferSent by examining if they are ecologically valid in its "learning environment" (i.e., the training set). In the following sections, we delve into the heuristic strategies that explain performance on our Comparisons dataset, as well as how InferSent might have come to encode them, by testing their ecological validity in the SNLI training dataset.

4.2.1. Overlap heuristic

We note in Fig. 3 that almost all the sentence pairs in the same-type comparisons were classified as entailments, despite half of them being true contradictions. A distinguishing feature of the same-type comparisons is that the premise and hypothesis sentences have full word overlap (they both contain exactly the same words). This observation allows us to hypothesize an *overlap heuristic*: High overlap in words between premise and hypothesis biases InferSent against classifying the pair as a contradiction.

While we have seen some evidence that this heuristic is indeed at play (based on the performance on the same-type comparisons), the question remains as to why it encodes this rule. With our knowledge of language, we know this simple rule to reflect on incorrect understanding of NLI. However, all the knowledge about the NLI task that InferSent encodes is from its training dataset. If the dataset has underlying structural regularities that can be exploited by simple heuristic strategies, then a tabula rasa model for NLI such as InferSent that is trained on this dataset will learn to encode it.

We carried out an analysis of the SNLI dataset to determine if the overlap heuristic is ecologically valid in it. First, we observed anecdotally that indeed several contradictory sentence pairs have relatively little overlap in words. For example, a contradictory sentence pair in SNLI is:

Premise: Several people are trying to climb a ladder in a tree.

Hypothesis: People are watching a ball game.

CONTRADICTION

To quantitatively verify this observation, we ranked all the sentence pairs in SNLI by overlap rate: $\frac{\#\text{ of overlap words}}{\text{total }\#\text{ of words}}$(in non-increasing order). We then considered the top X sentences with highest overlap for different Xs. As shown in Table 3, when considering the full dataset, the distribution is balanced (the percentage of entailments, contradictions, and neutral sentences are equal). However, we found that as the word overlap in the sentences increases, the percentage of contradictions drops. When considering only the top 1,000 sentence pairs for overlap, we found that 91.5% of them have entailment or neutral labels, with only the remaining 8.5% having a contradiction label.

It is therefore natural that InferSent encodes the simple overlap heuristic as a predictor against contradiction. This explains not only the failure of InferSent to generalize its good performance on SNLI to the same-type comparisons in our test dataset, but also matches the specific failure mode we observe in its responses.

4.2.2. Antonyms heuristic

We note in Fig. 3 the opposite trend for the more/less-type comparisons, where almost all the sentence pairs were classified as contradictions, despite half of them being true entailments. A distinguishing feature of the more/less-type comparisons is that the premise and hypothesis always differ by one word—if the premise contains the word “more” (“less”), then the hypothesis always contain the word “less” (“more”). This observation allows us to hypothesize an *antonyms heuristic*: Sentences differing in the presence of words that have opposing meanings (antonyms) tend to be classified by InferSent as contradictions, irrespective of the other words or their order in the sentence.

Similar to the previous section, we investigated the training dataset to elucidate if this heuristic is ecologically valid in InferSent’s training set. Anecdotally, we saw that the contradicting hypotheses provided by crowd workers to generate SNLI do follow this pattern. For example, a contradictory sentence pair in SNLI is:

Premise: A man in a white t-shirt takes a picture in the middle of the street with two public buses in the background.

Hypothesis: A man is wearing a black t-shirt.

CONTRADICTION

To verify this observation quantitatively, we analyzed the statistics of antonym usage in SNLI. To test whether a sentence pair (A,B) contains antonyms, we went through each word in sentence A, and considered all synonyms of that word, and considered all antonyms of those synonyms. Finally, we checked if sentence B contained any of those antonyms. These synonyms and antonyms were found using the NLTK WordNet software (Bird & Loper, 2004). We then considered two different statistics. First, we calculated P (Contradiction | Antonym), which is the probability that a sentence pair is a contradiction given that its premise and hypothesis contain an antonym pair. This measures how well the presence of antonyms predicts a contradiction label in the training set. Second, we

Table 3
Percentage of entailments split by overlap rate of words in SNLI

Top	Entailment	Neutral	Contradiction
All	33.4	33.3	33.3
10,000	39.5	35.7	24.8
1,000	50.8	40.7	8.5

calculated P(Antonym | Contradiction), which is the probability that a contradictory sentence pair contains antonyms. This measures how well a contradiction label predicts antonyms. Both statistics were compared with the equivalent statistic for entailment, to provide a baseline for comparison. Table 4 shows that the presence of antonyms strongly predicts a contradiction label in the SNLI dataset (61.2% compared to chance at 33.3%). We also found that a contradiction label predicts the presence of an antonym pair (12.2%) more strongly than entailment did (3.5%). This indicates that the antonyms heuristic can explain significant variance for the contradiction label in the training set.

Since most of our Comparisons dataset contained a large amount of overlap between premise and hypothesis, the rules InferSent applies when responding to these test questions might be biased toward those learned in similar high overlap settings during training. We checked the statistics of antonymy in the high overlap subset of SNLI (top 10,000 highest overlap) to provide a closer comparison (Table 5). Here, contradiction predicts the presence of an antonym pair (43.7%) more strongly than in the whole dataset (12.2%). The difference between P(Antonym | Contradiction) and P(Antonym | Entailment) is also more pronounced in this high overlap subset. The presence of an antonym pair no longer predicts contradictions at a high rate (28.9%), but this is possibly due to the very low base rate of contradictions in the high overlap subset of SNLI, as compared to entailments.

These results suggest again, that the underlying statistics of the SNLI dataset allow models, including InferSent, to perform well with simple lexical heuristics that ignore the order of words and their relations.

4.2.3. Negation heuristic

We see in Fig. 3 that the not-type comparisons are preferentially classified as contradictions. A distinguishing feature of the not-type comparisons is that the premise and the hypothesis differ by the presence of the negation “not.” This observation allows us to hypothesize a *negation heuristic* where sentence pairs that differ in the presence of negations are preferentially classified as contradictions.

Following procedures analogous to previous sections, we first noted anecdotally, that this heuristic seems to have validity in the contradicting hypotheses in SNLI. For example, a contradictory sentence pair in SNLI is:

Premise: Men turn to the camera to smile on the middle of three long tables in a refectory.

Hypothesis: The man is not smiling.

CONTRADICTION

Table 4
Percentage of entailments split by antonym word pair in the SNLI dataset

	P (Antonym | X)	P (X | Antonym)
X = Contradiction	12.2%	61.2%
X = Entailment	3.5%	18.0%

Table 5
Percentage of entailments split by antonyms in high overlap SNLI subset

	P (Antonym ∣ X)	P (X ∣ Antonym)
X = Contradiction	43.5%	28.9%
X = Entailment	8.7%	34.3%

We verified this observation quantitatively by looking at the statistics for negation in SNLI. We collected all sentence pairs that contain “negating N-grams”: no, not, n’t (by considering “n’t,” we included words such as “don’t” or “doesn’t”). We then carried out analyses similar to the previous section, where we checked (a) the predictive power of negations on contradictions (P(Contradiction ∣ Negation)), and (b) the predictive power of contradiction on negations, P(Negation ∣ Contradiction), and compare both of these to statistics for entailment as a baseline. We found (Table 6) that the presence of a negation strongly predicts contradiction in the SNLI dataset (58.4% compared to chance at 33.3%). We also found that while both numbers are very low, a contradiction predicts the presence of a negation (3.3%) slightly more strongly than entailment does (1.1%). We also carried out the same analysis for a high overlap subset (top 10,000 highest overlap) of SNLI to maximize similarity with our Comparisons dataset and saw similar results (Table 7). In fact, the presence of negation predicts a contradiction, P(Contradiction ∣ Negation) = 60.0%, at rates comparable to that in the full dataset, P(Negation ∣ Contradiction) = 58.4%, despite the much lower base rates of contradiction in this subset of the data. This indicates strong ecological validity for this heuristic in the high overlap subset of the SNLI dataset.

4.3. Summary of heuristics

We found evidence for three heuristics that explain the bulk of the patterns seen in the performance of InferSent on our Comparisons dataset, all of which are ecologically valid in the SNLI dataset. First, we identified the *overlap heuristic* where a large overlap in words between two sentences leads InferSent to not classify them as contradictions. Second, we identified the *antonyms heuristic* and the *negation heuristic*, where the premise and hypothesis differ in the presence of an antonym or a negation, which leads InferSent to classify them as contradictions.

These illustrate a disproportionate dependence on lexical (rather than relational) meaning in the representations and decision rules used by InferSent. While these heuristics serve well in certain domains, for example in SNLI, they do not amount to a more general encoding of entailment and contradiction between sentence pairs, as evidenced by InferSent’s poor performance on our Comparisons dataset.

The analysis so far has highlighted word-level heuristics that InferSent might be using. Yet the confusion matrix results (Fig. 3) show a slight asymmetry, indicating at least minor multi-word effects. This suggests that InferSent might be using some (potentially also heuristic) encodings for word order. However, a systematic analysis of the effect of

Table 6
Percentage of entailments split by negation in SNLI dataset

	P (Negation | X)	P (X | Negation)
X = Contradiction	3.3	58.4
X = Entailment	1.1	20.0

Table 7
Percentage of entailments split by negation in high overlap SNLI subset

	P (Negation | X)	P (X | Negation)
X = Contradiction	1.3	60.0
X = Entailment	0.1	7.5

word order, and how much variance such heuristics might explain, is challenging due to the combinatorial explosion in the number of possibilities. We leave a thorough investigation of this to future work.

5. Augmenting the learning environment

The foregoing results suggest that such ecological validity of simple heuristics in the SNLI training data (InferSent's learning environment) could explain why InferSent acquires them over a more abstract, systematic representation of the relations between words in a sentence. This leaves open the question of whether architectures such as InferSent are capable of learning the abstract relational rules needed to succeed at our task given a different training set where simple heuristics no longer explain so much of the variance. RNN architectures like the one in InferSent can in theory represent the relational structure required to encode the abstract rules of the sort in Table 1 (for details, see Section 2.3). But how might we get them to learn and use them? In this section, we explore this question by training the InferSent model on part of the Comparisons dataset, and testing on a held-out subset of it. This serves to test whether simple training on examples of the rules in Table 1 will enable InferSent to encode some abstract relational rules.

The total training subset of our Comparisons dataset consists of 40k sentence pairs (7% the size of the 550k pair SNLI training set). Validation and test sets consist of 2,000 sentence pairs each. There are no overlapping sentence pairs between any of these sets; therefore, simply memorizing the training set will not allow good test performance. Good test performance requires the encoding and utilization of an abstract relational rule.

We started with the original InferSent embeddings already trained on the SNLI dataset, and then fine-tuned these by training them on our new Comparisons dataset (using the same protocols used in Conneau et al., 2017, to train InferSent). Results are shown in Table 8. We found that using this method, performance on the SNLI data task degrades over the course of fine-tuning on the new Comparisons dataset from 84.84% to 56.37%. This points to over-fitting to the Comparisons data, at the cost of representing information necessary for SNLI. We found, however, that performance on the Comparisons test set is much higher (99.8%) than when trained only on SNLI (47.81%). Note that this test set consists of sentence pairs InferSent has never seen before. We thus find that the model architecture for InferSent, given the right training data, can encode some form of abstract relational structure that allows it to learn rules of the form in Table 1 and apply them to new sentence pairs—in particular sentence pairs with Xs, Ys, and Zs that it has never seen in that combination before.

Sequential training on different kinds of inputs (like the fine-tuning procedure above) is known to induce catastrophic forgetting in neural-network models (French, 1999), where solutions to previous tasks are overwritten by solution to new tasks. One possible remedy is to interleave training rather than train sequentially on different kinds of inputs (McClelland, McNaughton, & O'Reilly, 1995; for other approaches to combat catastrophic forgetting, see also Kirkpatrick et al., 2017; McRae and Hetherington, 1993). To check whether InferSent can represent this relational structure without losing the information necessary for SNLI, we started with an untrained network, and then trained on an augmented version of the original training data. Here, examples from the SNLI training set were randomly interleaved with examples from our Comparisons training dataset (by choosing examples uniformly at random across the joint training set consisting of both datasets), otherwise using the same training protocols reported in Conneau et al. (2017). The test results are reported in Table 9. We found that the accuracy obtained this way on the SNLI test set (84.96%) is comparable to the model trained only on SNLI (84.84%). Moreover, test accuracy on the Comparisons dataset is close to perfect (99.55%) and is much higher than the model trained only on SNLI (47.81%). This establishes that in this case the model has enough capacity to achieve high performance on specially designed edge cases like the Comparisons dataset, without loss of performance on the more general SNLI dataset.

Finally, we wanted to investigate the number of sentence pairs from the Comparisons dataset that would be required in the training set to achieve good performance on the Comparisons test set. The full size of the training set from our Comparisons dataset consists of 40k sentence pairs. In the experiments above, we include this entire dataset when

Table 8
Results of fine-tuning InferSent on the Comparisons dataset

Epoch	Performance (%)		
	Train(Comp)	Test(Comp)	Test(SNLI)
0	47.81	45.36	84.84
13	99.91	99.8	56.37

training InferSent; we now vary the percentage of this full training set that is included during training. The results from this are presented in Table 10. The training accuracy of all these runs (on the combined training data, including SNLI and varying numbers of Comparisons pairs) is around 90%. We see that including only 1k pairs from the Comparisons training set improves performance on the Comparisons test set slightly (from 45.36% to 58.91%), but still gives very poor performance overall. This indicates that with only 1k examples in its training set, the model does not generalize well to new combinations of nouns. However, performance steadily increases as the number of Comparison examples in the training set increase, hitting close to ceiling performance at about 10k sentence pairs. The final 30k sentence pairs (going from 10 to 40k examples in the training set) increase the test performance on the Comparisons test set, from 98.7% to 100.00%. Note that despite the number of training examples increasing, we keep the test set constant, and it is constructed such that the model has never seen the sentences in this test set during training.

This result also verifies that the heuristics we find in the original InferSent are an ecologically rational response to a training environment that licenses these “shortcut” strategies, and not because of shortcomings in representational or learning abilities of the model itself. This points to the benefits of understanding the learning environment in greater detail, and potentially including specially designed data to guard against incorrect heuristics that do not generalize. Research on the generation of adversarial examples targets this intuition. The idea is to have a separate “adversarial” model that generates edge-case training examples optimized to try and fool the main model into giving the wrong answer (Goodfellow, Shlens, & Szegedy, 2014; Zhao, Dua, & Singh, 2017). It does so by generating examples that violate the heuristics the main model has learned from training thus far. Subsequently, the training environment for the model is augmented to include these edge cases, making the current heuristics no longer ecologically valid. The main model therefore updates its representations and decision rules accordingly and the process is continued. Our work provides some insight into how we can leverage a top-down understanding of the structure of language and systematic stimulus design, to generate such edge-case training data and potentially improve the representations learned by machine-learning systems.

Table 9
Results of retraining Infersent on both SNLI and the Comparisons dataset

Epoch	Performance (%)		
	Train (Combined)	Test (Comp)	Test (SNLI)
0	33.21	0.00[a]	32.77
12	90.99	99.55	84.96

[a]The untrained InferSent model (using the default initialization procedure in PyTorch) classified almost all sentence pairs as neutral. This gives rise to chance accuracy on SNLI since roughly 33% of these examples are true neutral, and 0% accuracy on Comparisons because there are no true neutral pairs.

Table 10
Results of training Infersent on SNLI and varying number of pairs from the Comparisons dataset

Comp Pairs in Training	Performance (%)	
	Test(Comp)	Test(SNLI)
0	45.36	84.84
1,000	58.91	84.41
5,000	86.52	84.03
10,000	98.7	84.34
40,000	100.00	84.96

A key hurdle for the scalability for such augmentation as a solution to improving artificial representations of language, however, is that there are an infinite number of possible stimuli, with brand new combinations of words that may never have been encountered before. No finite amount of augmentation will allow a system to represent and process this infinite space of natural language sentences unless it can also *generalize* its knowledge gained from the examples observed thus far to new examples. In this section we saw that InferSent can generalize rules like those in Table 1 to never previously observed combinations of X, Y, and Z to perform well on the test set of the Comparisons dataset. In the following sections we further discuss the generalization capacities of the representations learned by InferSent, and we focus in particular on their differences and similarities to human generalization.

6. Generalization

An important and well-studied aspect of human-like representations is that rules learned with one set of tokens can be systematically generalized to other tokens (Fodor & Pylyshyn, 1988; Lake et al., 2019). In Section 6.1 we study if our machine-learned representations can perform such generalization to tokens that have never previously been observed. More often, however, the tokens to which we want to generalize learned rules have previously been observed, but simply in a different context. The historical contexts of tokens can determine some of their properties—like syntactic category and semantic content—which in turn inform how humans generalize rules to them, sometimes deviating from entirely systematic generalization. In Section 6.2 we examine how the historical context of tokens influences systematic generalization in our machine-learned representations, and how these effects compare to those in humans.

Throughout this section, we will only consider sentence pairs that are similar in structure to ones in our Comparisons dataset, and we will no longer consider performance on SNLI. We will predominantly be studying the model that has been trained jointly with our Comparisons dataset in addition to SNLI (referred henceforth to as the augmented-InferSent model).

6.1. Zero-shot reasoning

Zero-shot reasoning is the ability to solve tasks involving a term that has never been seen before. This (often also called zero-shot learning) has commonly been used as a test for systematicity (Lake et al., 2017)—a human can carry out inferences like "Anne is more boffy than Bob" entails that "Bob is less boffy than Anne" without ever having encountered the word "boffy" before.

But this ability requires the representation learned to be abstract, and not be tied to the Xs, Ys, and Zs seen in training. Instead, it has to encode an abstract relational rule where "X is more Y than Z" entails "Z is less Y than X" for all possible X, Y, and Z, irrespective of their specific values. If the representations are tied to the observed values of Xs, Ys, and Zs and cannot generalize to new values for these, each possible X, Y, and Z has to have occurred in the training dataset. However, these can be arbitrarily complex (e.g., "The old woman with a flower in her hair *is more deliriously happy than* the tall young man wearing the blue bowler hat" implies that "The tall young man wearing the blue bowler hat *is less deliriously happy than* the old woman with a flower in her hair"). Ensuring that every possible such X, Y, and Z have been seen in the training data is impossible, and this kind of generalization is key to human-like language understanding.

In this section we consider the performance of the augmented-InferSent model. We already know that this model performs well on SNLI, and generalizes to new combinations of X, Y, and Z in our Comparisons dataset (see Table 9), where each X, Y, and Z have previously been seen. In this section, we analyze its ability to generalize to three different kinds of Xs and Zs that have never been encountered during training.

1. **Held out nouns**: Nouns (from the GloVe dataset) that never occur in the training data (neither SNLI nor our Comparisons dataset).
2. **Made up "words"**: Directly using a 300 dimensional vector randomly sampled from an uncorrelated Gaussian distribution, as a stand-in for a real GloVe vector.
3. **Long noun phrases**: The Xs and Zs used in training as part of the Comparisons dataset were of the type "the man." Here we generate longer noun phrases of the form "the grumpy man in front of us" consisting of randomly sampled adjectives, nouns, and prepositional phrases.
4. **Very long noun phrases**: To test how far our system generalizes, we also test on extremely long noun phrases. Here we chain four adjectives to the front of the noun and add a modifier phrases to the end. An example is: "the tall funny agreeable old man sitting on the chair." If the system is abstracting "rules" at the level of noun phrases, it should succeed at this test, but if the abstraction is based on single words, this test should be more difficult.

For each subtype in the Comparisons dataset (same, more-less, and not types), we generated a test set of 1,000 sequences by substituting Xs and Zs of the above kinds. The Ys were sampled in the same way as in the Comparisons dataset (random adjectives that appear in SNLI). We then tested on these sentences and reported the average accuracy. Note that not only had these specific sentences (combinations of X, Y, and Z) never been seen during training, even the individual Xs and Zs had not been seen. We found that

InferSent generalizes to all three new kinds of Xs and Zs quite well (Table 11). The held-out nouns are the most similar to the Xs and Zs seen during training since they are also exactly one word and are nouns sampled from GloVe. It is notable that generalization performance with these is comparable to that with the very different kinds of Xs and Zs such as the made-up words, or longer noun phrases, indicating a fairly abstract representation of relational rules that are not tied to the specific value of X and Z. However, generalization performance for the very long noun phrases, which are similar but require greater long-distance coherence, drops significantly, indicating that this zero-shot generalization does not abstract over complex noun phrases in general.

These results indicate that the representation learned by augmented-InferSent is partially abstract and composable, allowing some degree of systematic generalization to a variety of Xs and Zs that have never been seen before. In the next section we further probe contextuality of generalization and how that interacts with the training set/learning environment, making comparisons to human generalization.

6.2. Context-tying

We saw in the previous section that augmented-InferSent has some of the central human-like capacity of zero-shot reasoning. This indicates some systematicity in its representations. However, even humans do not always succeed at fully systematic generalization. In this section we investigate these exceptions and qualifications to the widest interpretation of systematic generalization, focusing on the role of context in generalization. We do this in two ways: using type violations and biased exposure.

6.2.1. Type violations

One extreme of learning a purely abstract rule like in Table 1 is to be completely insensitive to any properties of the Xs, Ys, and Zs, and generalize this rule to all possible tokens. However, this very strong generalization may not always match human intuitions. For example, the sentence pair

Premise: The punctual is more cheerful than the man.

Hypothesis: The punctual is not more cheerful than the man.

does not seem to have a right answer. The rule applies easily only to Xs, Ys, and Zs that are of the right type—in this case the right syntactic category.

Table 11
Zero-shot reasoning: Performance on previously unobserved Xs and Zs

Test Set	InferSent (%)	Augmented-InferSent (%)
Held-out nouns	47.9	82.0
Made up words	48.0	83.2
Long noun phrases	49.1	84.9
Very long noun phrases	49.7	59.3

While syntactic structure is not directly provided to the embedding model, some notion of syntactic category will be implicit. Information about the syntactic category of a word can be gleaned from its contexts, that is, the other words around it (Chomsky, 1993; Redington et al., 1998; Socher, Manning, & Ng, 2010), and in some cases can be decoded from word embeddings directly (Pennington et al., 2014).

We investigated generalization of rules in augmented-Infersent to test items which, unlike in the previous section, had been previously seen, but had only occurred in a different syntactic role (i.e., a different context). We generated a test set of ungrammatical sentences using Xs and Zs that are random non-nouns, in our case random adjectives from SNLI. Crucially, these words had been seen before, but never in the position/context that X and Z occupy in the Comparisons dataset, since appearing in those positions violates syntax. We then evaluated the performance of the augmented-InferSent model in the same way as in the previous section on zero-shot generalization. We found that accuracy on such sentence pairs is low, giving poor performance (Table 12). This indicates that the rules learned, though at least partially abstract as indicated by generalization to held-out nouns, come with restrictions on the type of (known) items they will apply to. This follows closely how humans generalize—that learned rules do not generalize indiscriminately to all tokens, but rather only within some fixed categories. These categories in turn, like syntactic categories, can be gleaned from the contexts in which these tokens usually appear. In the next section, we examine the role of semantic content in the context of tokens, and how that influences generalization.

6.2.2. Biased exposure

In this section, we manipulate the context of various tokens, without violating the syntactic rules, to study its effect on generalization. In all the augmentations we have used so far, some token X is equally likely to occur in the context of a same-type sentence pair as it is in the context of a more/less-type sentence pair. Similarly, X is as likely to occur in the context where it is "more cheerful than the man" as it is to occur in the context where it is "less cheerful than the man." Therefore, aside from the restrictions of syntactically correct placement, there is no additional structure around which contexts which tokens occur in—they are all randomly distributed. However, in the real world, tokens are not uniformly sampled into contexts even within constraints of syntax; a word is much more likely to be sampled repeatedly in certain contexts than others. This is because the appearance of tokens in naturally occurring sentences is not determined solely by their syntactic role, but also by their semantic role. For example, one is much more likely to encounter the sentence "Broccoli is more nutritious than candy" than the sentence "Candy is more nutritious than broccoli," since one is true of the real world, and the other is not. Nonetheless, the premise "Candy is more nutritious than broccoli" still logically entails the hypothesis "Broccoli is less nutritious than candy." Statistics of how often certain implications and inferences are made in the learning environment (that will be reflected in semantic beliefs about the real world) can interfere with such logical inferences in humans in both deductive (Evans, Barston, & Pollard, 1983) and

Table 12
Type violations: Performance with tokens from the wrong syntactic category, versus with held-out tokens from the right syntactic category

Test Set	InferSent (%)	augmented-InferSent (%)
Held-out nouns	47.9	82.0
Non-noun words	47.9	49.3

probabilistic (Evans, Handley, Over, & Perham, 2002) reasoning. This is often termed "belief bias."

In this section, we test if the representations we are studying exhibit belief bias. We manipulate the uniformity in the co-occurrence of tokens with contexts (subject to syntactic constraints), and we examine if a newly augmented InferSent model can generalize a token it has seen in one context, to cases where it appears in a different context. We compare this to a zero-shot control condition, where the test token has never been seen before.

To this end, we first generated variants of our Comparisons dataset where tokens are no longer uniformly sampled into contexts. We considered only two subtypes of the comparison types summarized in Table 1: the *same-type* (C_2) and the *more/less-type* (C_1). These constitute the two contexts C_1 and C_2 in which tokens can appear. Noun phrases were generated using the same procedure used for the long noun phrases in the section on zero-shot reasoning—phrases (tokens) of the form "The grumpy man in front of us." These tokens were then randomly divided into T^0-type and T^*-type (460 each). Therefore, there is no structural difference between the T^0 and T^* tokens, only the context in which they are seen will differ across conditions.

We built four sets of sentence pairs that vary in their context–token combination: C_2T^0 consisted of combinations of T^0 tokens in a C_2 context, so on and so forth for C_2T^*, C_1T^0, and C_1T^*. Each such context–token combination set was independently divided into train and test sets (each of size 5,000). The sentence pairs in each of the four test sets had never been seen before in any of the four training sets.

We augmented the original InferSent embeddings with different combinations of samples from the four different train sets.[3] We then compared their performance on all four of the test sets to examine how different context–token combinations seen during training influenced test generalization. The three different embeddings that result are as follows:

1. **Zero-shot control condition**: Only the T^0 tokens were seen in training; no T^* token were seen at all. Therefore testing with tokens from T^* is analogous to zero-shot reasoning. The training set consisted of the full training sets from C_1T^0 and C_2T^0.
2. **Experimental conditions**: Both T^0 and T^* tokens were seen in training, therefore testing with tokens from T^* is not analogous to zero-shot reasoning. However, the contexts in which T^0 and T^* tokens appear during training differed. There are two different embeddings we trained of this kind.

a. Exposed—$\boldsymbol{C_1T^*}$: This embedding saw T^0 tokens in both C_1 and C_2 contexts (as with the control condition), and additionally also saw T^* tokens—but only in the C_1 context. In order to balance the number of training examples from each context between conditions, the training set consisted of the full training sets from C_2T^0 and half (randomly selected) of the training set from each of the C_1T^0 and C_1T^* context–token combination sets.
b. Exposed—$\boldsymbol{C_2T^*}$: This embedding saw T^0 tokens in both C_1 and C_2 contexts, but saw T^* tokens only in the C_2 context. The training set was balanced across contexts here as well.

All three models received the same number of training examples, with equal numbers of sentence pairs from both contexts C_1 and C_2. They all also saw T^0 noun phrases appear in both contexts. The three models only differed in which contexts T^* noun phrases appeared during training. The control model never saw T^* noun phrases, Exposed—C_1T^* only saw them in the C_1 context and Exposed—C_2T^* only saw them in the C_2 context. All of these were then tested on the same held-out test set. We see from Table 13 that all three models generalize well to held-out test examples involving previously unobserved combinations of T^0 noun phrases in both contexts (first row). This is consistent with our initial results on augmentation (see Section 5). Further, the control (zero-shot reasoning) condition that never saw noun phrases in training generalizses well to all the test examples with noun phrases (first column). This is consistent with our results on zero-shot generalization (see Section 6.1).

We now turn to generalization performance when tokens were seen before but only in a specific context (second and third columns in Table 13). We discuss the results for the model Exposed—C_1T^* (that saw T^* noun phrases in C_1 type comparisons), a symmetric discussion applies also to Exposed—C_2T^*. We found that Exposed—C_1T^* performs well on held-out test examples from the C_1T^* category (99.7%)—as consistent with our original experiments with augmentation. However, we found that it fails to generalize very well to T^* type noun phrases in the C_2 context, with a significant drop in performance (67.71%). The crucial comparison is that this low performance is also significantly worse than that of the zero-shot control on the same test set (95.78%). Neither of these has seen T^* phrases in the C_2 context—yet the control generalizes very well, while the Exposed—C_1T^* fails to. This indicates that while the representations learned can generalize well to previously unseen tokens, this generalization is poorer to tokens that have in fact been seen before, but only in a different context.

This indicates that our representations do learn something akin to belief bias, where the context in which tokens have been seen (even within the right syntactic category) can influence how abstract logical rules (like in Table 1) generalize to them. This suggests potential directions for research on modeling how belief bias in humans arises. However, it is crucial to point out that although humans do exhibit such context-tying, the effects are mostly observed in children (Evans & Perry, 1995) and under time pressure/cognitive load (Evans & Curtis-Holmes, 2005). The coexistence of such a fast heuristic strategy (that potentially suffers from belief bias), and a slower deliberative strategy (that can

Table 13
Biased exposure: Results from InferSent embeddings augmented with different training sets that manipulate the co-occurrence of context and token

Test Set	Performance (%)		
	Zero-Shot	Exposed–C_1T^*	Exposed–C_2T^*
$C_1T^0 + C_2T^0$	97.44	97.02	98.0
C_1T^*	95.72	99.7	61.16
C_2T^*	95.78	67.71	99.96

perform abstract reasoning) is a well-studied and popular model for representations and decision rules in humans (Evans & Curtis-Holmes, 2005; Groves & Thompson, 1970; Kahneman, 2011). Thus, although people have a tendency toward belief bias, they are able to overcome it and engage in abstract reasoning, which our machine-learned representations cannot do.

This raises a new concern about the scalability of augmentation as a general approach to learning systematic representations in such tabula rasa machine-learning systems. There are infinitely many possible sentences that all follow the rules of syntax, so observing tokens in contexts that one has not often seen them in, but where they are syntactically valid, is likely to occur often. Our new findings show that while zero-shot reasoning to previously unobserved tokens works in certain cases, these tabula rasa systems may tie an observed token to the small fraction of contexts in which it has been seen. This hinders generalization to cases where this token occurs in a new context. In order for every token to have been observed in every context, a combinatorially large amount of augmented training data would be required, potentially making this approach unfeasible for achieving the kinds of systematic representations humans have.

7. Discussion and future work

In this paper, we carried out a case study in the use of methods from cognitive science and psycholinguistics to better understand machine-learned representations. We developed minimal cases in an NLI task that test for some aspects of abstract relational structure in sentences. We used this diagnostic tool on a large-scale state-of-the-art NLI model (Conneau et al., 2017) to not only demonstrate that there is no evidence of abstract composable structure in the sentence embeddings, but also provide insight into the representations and decision criteria actually learned. This approach led us to isolate the use of some simple heuristics, which we then traced to structural regularities in the training distribution. This allowed us to demonstrate the strong effect the training data have on the representations learned. We then augmented this training environment with so-called adversarial examples such that simple heuristics like the ones we found are no longer ecologically valid. We found that such augmentation is one way to lead the system to learn some forms of abstract relational structure. Notably, we found that one of the traditional holy grails of systematicity—zero-shot generalization of learned rules to new, previously unseen words—can be partially achieved with this method. Further tests, however, revealed limitations in the breadth of this generalization. We found that while zero-shot generalization to previously unseen words works, generalizations to words that have previously been seen in a different context suffer. This gives us another measure for the extent of systematicity in representations—a phenomenon we call "context-tying." We discussed the relationship between this effect and findings in human cognitive psychology where semantic beliefs about the real world can interfere with flexible inferences supported by abstract logical representations (Evans, 2013). This parallel suggests new ways to model this psychological phenomenon (Dasgupta, Schulz, Tenenbaum, & Gershman, 2020). The presence of context-tying in the machine-learned representations indicates that these systems might have to see each word in a wide variety of contexts in order to generalize it to new contexts, indicating that combinatorially large amounts of augmentation will likely be required for a tabula rasa unstructured neural network model to learn an entirely systematic representation from data.

These results suggest many directions for future work. We showed how the issue of context-tying bodes poorly for the scalability of using only training data augmentations to achieve human-like systematic representations. Recent work, however, suggests such adversarial mechanisms in the human brain (Gershman, 2019). This motivates further research on how this approach might be made more scalable. We studied the representations learned from a relatively fixed amount of augmentation and training. An important step forward is to better understand how systematicity in these representations evolves over the course of augmented training, and exactly how much augmentation is really needed. Another important problem is to understand what augmentations work best. To that end, a promising direction is to integrate our approach (where augmentations are generated using existing knowledge about analogous representations in humans) with approaches that learn to generate such adversarial augmentations (Goodfellow et al., 2014; Kang et al., 2018; Zhao et al., 2017). Finally, we did not examine the sentence representations directly, but rather via the performance of a downstream classifier trained end-to-end with the sentence embeddings. Future work should also consider other ways to directly analyze these representations.

Human infants are not as tabula rasa as models like InferSent but rather encode useful inductive biases (Chomsky & Lightfoot, 2002; Lightfoot & Julia, 1984; Mitchell, 1980; Pearl & Goldwater, 2016; Seidenberg, 1997). Building such biases into our models (Battaglia et al., 2018; Dubey, Agrawal, Pathak, Griffiths, & Efros, 2018; Gandhi & Lake, 2019; Lake et al., 2017) is a promising direction toward scalably learning systematic representations. We also showed how analysis and controlled testing for heuristic strategies in the learning environment can provide rich insights into the representations learned. Such analyses could also be used to improve learning and subsequent performance by leveraging this underlying structure (Gigerenzer & Todd, 1999; Martignon & Hoffrage, 2002; Şimşek, 2013; Şimşek, Algorta, & Kothiyal, 2016). Finally, we leverage methods from cognitive psychology to introduce a new structured test dataset (the Comparisons dataset) as well as a new metric (context-tying) for sentence representations. This approach joins other recent approaches (Belinkov & Glass, 2019; Ettinger et al., 2018; McCoy et al., 2019) in going beyond the traditional single-dimensional metrics of the aggregate performance achieved on test datasets that match the distribution of training, and additionally provides insights into the specific kinds of mistakes made. This paves the path forward to more principled and nuanced ways to benchmark artificial systems against humans (Lake & Baroni, 2018; Linzen, Dupoux, & Goldberg, 2016). Further, a metric like context-tying is not bound to the domain of language, and can also be used to benchmark systematicity in other domains that benefit from abstract compositional representations—like scene understanding (Johnson et al., 2016; Ommer & Buhmann, 2009) or structured planning (Burridge, Rizzi, & Koditschek, 1999; Singh, 1992). Future work should pursue other such diagnostic metrics, to build toward a comprehensive suite of testable criteria for exactly what constitutes human-like representations, and also to further inform which aspects of these we wish to emulate in artificial systems.

Acknowledgments

We are grateful to Anatole Gershman, Tim O'Donnell, Joshua Greene, and Fiery Cushman for helpful discussions. I.D. is supported by Microsoft Research. S.J.G. is supported by the Office of Naval Research (N00014-17-1-2984) and the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. N.D.G. is supported by DARPA agreement number FA8750-14-2-0009, and a Sloan Foundation Research Fellowship. Code and data are available at github.com/ishita-dg/ScrambleTests. A preliminary version of this work was previously reported in Dasgupta, Guo, Stuhlmüller, Gershman, and Goodman (2018).

Notes

1. The details and implementation of the optimization algorithm also contribute (for an overview, see Ruder, 2016), but as long as the optimization reaches convergence this has relatively little effect, and we leave this out of our current discussion.
2. Only a few words differed by more than 1% from their occurrence rate in SNLI, such as *not*, *a*, *than*, *the*, *is*, *less*, *more*. This was inevitable given the general structure of the comparison sentence pairs we use. All of these words however did still occur in the SNLI training corpus, and were not new to the model at test time.
3. In this experiment we only make comparisons between the performances of differently augmented models, rather than considering the overall performance like in previous experiments. The influence on performance from the SNLI training data is irrelevant since it will affect all four augmented models equally. Therefore, we can neglect SNLI performance and carry out our experiments using fine-tuned augmentation rather than full retraining (for details on these, see Section 5). This is computationally a lot cheaper.

References

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., Dyer, C., Heess, N., Wierstra, D., Kohli, P., Botvinick, M., Vinyals, O., Li, Y., & Pascanu, R. (2018). Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:1806.01261.

Beals, R., Krantz, D. H., & Tversky, A. (1968). Foundations of multidimensional scaling. *Psychological Review*, *75*(2), 127.

Belinkov, Y., & Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, *7*, 49–72.

Bentivogli, L., Bernardi, R., Marelli, M., Menini, S., Baroni, M., & Zamparelli, R. (2016). Sick through the semeval glasses. lesson learned from the evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *Language Resources and Evaluation*, *50*(1), 95–124.

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, *124*(1), 1–20.

Bird, S., & Loper, E. (2004). Nltk: The natural language toolkit. Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, 31.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326.

Braine, M. D. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review*, *85*(1), 1.

Burridge, R. R., Rizzi, A. A., & Koditschek, D. E. (1999). Sequential composition of dynamically dexterous robot behaviors. *The International Journal of Robotics Research*, *18*(6), 534–555.

Chomsky, N. (1993). *Lectures on government and binding: The pisa lectures (No. 9)*. Berlin: Walter de Gruyter.

Chomsky, N., & Lightfoot, D. W. (2002). *Syntactic structures*. Berlin: Walter de Gruyter.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364.

Dasgupta, I., Guo, D., Stuhlmüller, A., Gershman, S. J., & Goodman, N. D. (2018). Evaluating compositionality in sentence embeddings. arXiv preprint arXiv:1802.04302.

Dasgupta, I., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2020). A theory of learning to infer. *Psychological Review*, *127*(3), 412.

Dubey, R., Agrawal, P., Pathak, D., Griffiths, T. L., & Efros, A. A. (2018). Investigating human priors for playing video games. arXiv preprint arXiv:1802.10217.

Ettinger, A., Elgohary, A., Phillips, C., & Resnik, P. (2018). Assessing composition in sentence vector representations. arXiv preprint arXiv:1809.03992.

Evans, J. S. B. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7* (10), 454–459.

Evans, J. S. B. (2013). *The psychology of deductive reasoning (psychology revivals)*. London: Psychology Press.

Evans, J. S. B., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, *11*(3), 295–306.

Evans, J. S. B., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning*, *11*(4), 382–389.

Evans, J. S. B., Handley, S. J., Over, D. E., & Perham, N. (2002). Background beliefs in Bayesian inference. *Memory & Cognition*, *30*(2), 179–190.

Evans, J. S. B. T., & Perry, T. S. (1995). Belief bias in children's reasoning. *Cahiers De Psychologie Cognitive/Current Psychology of Cognition*, *14*(1), 103–115.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3–71.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, *3*(4), 128–135.

Gandhi, K., & Lake, B. M. (2019). Mutual exclusivity as a challenge for neural networks. arXiv preprint arXiv:1906.10197.

Gershman, S. J. (2019). The generative adversarial brain. *Frontiers in Artificial Intelligence*, *2*, 18.

Gershman, S., & Tenenbaum, J. B. (2015). Phrase similarity in humans and machines. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th annual meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In *Simple heuristics that make us smart* (pp. 3–34). Oxford: Oxford University Press.

Glockner, M., Shwartz, V., & Goldberg, Y. (2018). Breaking nli systems with sentences that require simple lexical inferences. arXiv preprint arXiv:1805.02266.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Groves, P. M., & Thompson, R. F. (1970). Habituation: A dual-process theory. *Psychological Review*, *77*(5), 419.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. (2018). Annotation artifacts in natural language inference data. arXiv preprint arXiv:1803.02324. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Hill, F., Cho, K., & Korhonen, A. (2016). Learning distributed representations of sentences from unlabelled data. arXiv preprint arXiv:1602.03483. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*(2), 251–257.

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, *310*(5749), 863–866.

Jia, R., & Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. arXiv preprint arXiv:1707.07328.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., & Girshick, R. (2016). CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. arXiv preprint arXiv:1612.06890. Proceedings of the IEEE conference on computer vision and pattern recognition.

Johnson-Laird, P. N., & Steedman, M. (1978). The psychology of syllogisms. *Cognitive Psychology*, *10*(1), 64–99.

Jones, M. N., & Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1.

Kádár, A., Chrupała, G., & Alishahi, A. (2017). Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, *43*(4), 761–780.

Kahneman, D. (2011). *Thinking, fast and slow*. London: Macmillan.

Kang, D., Khot, T., Sabharwal, A., & Hovy, E. (2018). Adventure: Adversarial training for textual entailment with knowledge-guided examples. arXiv preprint arXiv:1805.04680.

Karpathy, A., Johnson, J., & Fei-Fei, L. (2015). Visualizing and understanding recurrent networks. arXiv preprint arXiv:1506.02078.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, *114*(13), 3521–3526.

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R., & Fidler, S. (2015). Skip-thought vectors. arXiv preprint arXiv:1506.06726.

Lai, A., & Hockenmaier, J. (2014). Illinois-lh: A denotational and distributional approach to semantics. SemEval@ COLING, 329–334.

Lake, B. M., & Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. arXiv preprint arXiv:1711.00350.

Lake, B. M., Linzen, T., & Baroni, M. (2019). Human few-shot learning of compositional instructions. arXiv preprint arXiv:1901.04587.

Lake, B., Ullman, T., Tenenbaum, J., & Gershman, S. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, E253. https://doi.org/10.1017/S0140525X16001837

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436.

Li, J., Chen, X., Hovy, E., & Jurafsky, D. (2015). Visualizing and understanding neural models in nlp. arXiv preprint arXiv:1506.0106.

Lightfoot, D., & Julia, P. (1984). *The language lottery: Toward a biology of grammars*. Cambridge, MA: MIT Press.

Linzen, T. (2019). What can linguistics and deep learning contribute to each other? Response to pater. *Language*, *95*(1), e99–e108. https://doi.org/10.1353/lan.2019.0001

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535.

Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

Martignon, L., & Hoffrage, U. (2002). Fast, frugal, and fit: Simple heuristics for paired comparison. *Theory and Decision*, *52*(1), 29–71.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419.

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. arXiv preprint arXiv:1902.01007.

McRae, K., & Hetherington, P. A. (1993). Catastrophic interference is eliminated in pretrained networks. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society: June 18 to 21, 1993, Institute of Cognitive Science, University of Colorado-Boulder* (pp. 723–728). London: Lawrence Erlbaum Associates.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111–3119.

Mitchell, T. M. (1980). *The need for biases in learning generalizations* (pp. 184–191). Piscataway, NJ: Department of Computer Science, Laboratory for Computer Science Research, Rutgers Univ.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529.

Nie, Y., Wang, Y., & Bansal, M. (2019). Analyzing compositionality-sensitivity of nli models. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, 6867–6874.

Ommer, B., & Buhmann, J. (2009). Learning the compositional nature of visual object categories for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(3), 501–516.

Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio (p. 125). 9th ISCA Speech Synthesis Workshop.

Pearl, L., & Goldwater, S. (2016). Statistical learning, inductive bias, and Bayesian inference in language acquisition. In J. L. Lidz, W. Snyder, & J. Pater (Eds.), *The Oxford handbook of developmental linguistics*. Oxford: Oxford University Press.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Doha: Association for Computational Linguistics.

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, *33*(3–4), 175–190.

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., & Van Durme, B. (2018). Hypothesis only baselines in natural language inference. In M. Nissim, J. Berant, & A. Lenci (Eds.), *Proceedings of the seventh joint conference on lexical and computational semantics*. New Orleans: Association for Computational Linguistics.

Redington, M., Crater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*(4), 425–469.

Ritter, S., Barrett, D. G., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. *International Conference on Machine Learning*, 2940–2949.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Rumelhart, D. E., & Abrahamson, A. A. (1973). A model for analogical reasoning. *Cognitive Psychology*, *5* (1), 1–28.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, Li (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Seidenberg, M. S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science*, *275*(5306), 1599–1603.

Siegelmann, H. T., & Sontag, E. D. (1995). On the computational power of neural nets. *Journal of Computer and System Sciences*, *50*(1), 132–150.

Simon, H. A. (1991). Bounded rationality and organizational learning. *Organization Science*, *2*(1), 125–134.

Simsek, Ö. (2013). Linear decision rule as aspiration for simple decision heuristics. *Advances in Neural Information Processing Systems*, *2*, 2904–2912.

Simsek, Ö., Algorta, S., & Kothiyal, A. (2016). Why most decisions are easy in tetris—And perhaps in other sequential decision problems, as well. *Proceedings of Machine Learning Research*, *48*, 1757–1765.

Singh, S. P. (1992). Transfer of learning by composing solutions of elemental sequential tasks. *Machine Learning*, *8*(3–4), 323–339.

Socher, R., Manning, C. D., & Ng, A. Y. (2010). Learning continuous phrase representations and syntactic parsing with recursive neural networks. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010* (pp. 1–9). Vancouver: Curran Associates, Inc.

Steyvers, M. (2006). Multidimensional scaling. In L. Nadel (Ed.), *Encyclopedia of cognitive science*. https://doi.org/10.1002/0470018860.s00585

Todd, P. M., & Gigerenzer, G. (2007). Environments that make us smart: Ecological rationality. *Current Directions in Psychological Science*, *16*(3), 167–171.

Tsuchiya, M. (2018). Performance impact caused by hidden bias of training data for recognizing textual entailment. arXiv preprint arXiv:1804.08117.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *European Conference on Computer Vision*, *8689*, 818–833.

Zhao, Z., Dua, D., & Singh, S. (2017). Generating natural adversarial examples. arXiv preprint arXiv:1710.11342.