ANNUAL REVIEWS

Annual Review of Linguistics

Semantic Structure in Deep Learning

Ellie Pavlick

Department of Computer Science, Brown University, Providence, Rhode Island, USA; email: ellie_pavlick@brown.edu

Annu. Rev. Linguist. 2022. 8:447–71

First published as a Review in Advance on November 23, 2021

The *Annual Review of Linguistics* is online at linguistics.annualreviews.org

https://doi.org/10.1146/annurev-linguistics-031120-122924

Keywords

deep learning, semantics, natural language processing, neural network interpretability, neural network analysis

Abstract

Deep learning has recently come to dominate computational linguistics, leading to claims of human-level performance in a range of language processing tasks. Like much previous computational work, deep learning–based linguistic representations adhere to the distributional meaning-in-use hypothesis, deriving semantic representations from word co-occurrence statistics. However, current deep learning methods entail fundamentally new models of lexical and compositional meaning that are ripe for theoretical analysis. Whereas traditional distributional semantics models take a bottom-up approach in which sentence meaning is characterized by explicit composition functions applied to word meanings, new approaches take a top-down approach in which sentence representations are treated as primary and representations of words and syntax are viewed as emergent. This article summarizes our current understanding of how well such representations capture lexical semantics, world knowledge, and composition. The goal is to foster increased collaboration on testing the implications of such representations as general-purpose models of semantics.

1. INTRODUCTION

Distributional semantics models (DSMs), which define the meanings of words in terms of the contexts in which they are used (Harris 1954, Firth 1957), have received significant attention in computational linguistics and natural language processing (NLP) (Erk 2012, Boleda 2020). Recently, traditional DSMs have fallen to the wayside in favor of linguistic representations derived from deep learning (LeCun et al. 2015). This shift has been associated with dramatic advances in the empirical performance of language processing systems, including claims of human performance on answering questions (Rajpurkar et al. 2016a) and engaging in dialogue (Adiwardana et al. 2020).

Like traditional DSMs, these new deep learning representations are derived from word co-occurrences in text corpora. However, unlike traditional DSMs, which construct sentence representations bottom-up via explicit composition of lexical representations, new deep learning models take a top-down approach in which the primary goal is to represent sentences in a way that supports commercial applications, and representations of words and their composition are viewed as emergent. Thus, while they build on the same fundamental idea of meaning-from-use as traditional DSMs, new deep learning representations entail fundamentally new models of lexical and compositional meaning, which are ripe for theoretical analysis.

So far, however, deep learning–based representations have not engaged rigorously with semantic theory, making it hard to differentiate genuine progress in modeling semantics from merely an increase in models' ability to memorize corpus statistics. This article summarizes our current understanding of how well representations derived from deep learning capture "semantics," broadly construed, with the goal of fostering increased collaboration on formalizing and testing the implications of such representations as general-purpose models of semantics. After some preliminaries (Sections 2 and 3), the review focuses on experiments in the areas of lexical semantics (Section 4.1), world knowledge (Section 4.2), and composition (Section 4.3), highlighting evidence for and against deep learning's ability to account for key phenomena within each.

2. DISTRIBUTIONAL SEMANTICS VERSUS DEEP LEARNING

This section briefly contrasts traditional DSMs with newer deep learning models. For discussions of DSMs in the context of linguistic theory, readers are referred to Lenci (2008, 2018), Boleda & Herbelot (2016), and Boleda (2020). For overviews of empirical methods and results, readers are referred to Turney & Pantel (2010) and Erk (2012).

2.1. Distributional Semantics Models

The distributional hypothesis (Harris 1954, Firth 1957) states that the meaning of a word is defined by the contexts in which it is used. In NLP, "contexts" is typically taken to mean linguistic contexts—that is, the other words with which a word co-occurs. Distributional semantics as a theory does not require this formulation; for instance, distributions could in principle be defined over representations of visual or emotional contexts instead. However, linguistic contexts are readily observable via text corpora, and modeling word meanings in this way has proven effective at explaining human similarity judgments (Landauer & Dumais 1997), priming effects (Lund et al. 1995), and selectional preferences (Erk 2007). Thus, in practice, many authors use the phrase "distributional semantics" to refer specifically to this formulation (Emerson 2020).

DSMs are almost always instantiated as vector space models—that is, models that represent words as points in space. Research on DSMs has focused on both lexical and compositional semantics.

Focusing on lexical representations, the simplest DSMs use count-based vectors and represent words as points in a sparse, very high-dimensional space. In such models, a word w_i in a vocabulary of size V can be represented by a V-dimensional vector in which the jth component reflects the number of times w_i occurs in the context of w_j. Typically, the context is a small window (e.g., ± 2 words), though many variations exist (Turney & Pantel 2010). An alternative to count-based DSMs is prediction-based DSMs, often called word embeddings, which represent words as dense, low-dimensional approximations of the explicit count distributions. Word embeddings are produced either by applying dimensionality reduction to the count matrices (Dumais et al. 1988, Landauer & Dumais 1997, Dhillon et al. 2011) or by using neural networks (Bengio et al. 2003, Mikolov et al. 2013, Pennington et al. 2014). Unlike count-based vectors, dimensions of word embeddings correspond to subsymbolic dimensions that yield good predictions and are not immediately interpretable. However, prediction-based embeddings have been associated with better explanatory power across many tasks (Baroni et al. 2014).

Work on compositional DSMs seeks to provide a story of how the above-described word representations combine to produce sentence representations. For example, Smolensky (1990), Baroni & Zamparelli (2010), and Grefenstette & Sadrzadeh (2011) use dimensionality to encode semantic type—for instance, representing nouns as N-dimensional vectors and adjectives as $N \times N$-dimensional matrices, thus mirroring the idea of predicate-argument structure mathematically. However, such approaches result in sentence representations that vary in size and shape based on their length and syntactic structure, and in practice, they are outperformed by simple composition methods (e.g., order-insensitive vector addition) (Mitchell & Lapata 2010, Hartung et al. 2017). As a result, compositional DSMs have not produced viable systems for processing language outside of a toy setting.

2.2. Sentence Encodings from Deep Learning

Recent years have seen a shift away from traditional DSMs in favor of using larger, more complex deep learning models. Unlike traditional DSMs, these newer models at least give the appearance of processing complex language (e.g., performing impressively well at real-world tasks like reading comprehension; Rajpurkar et al. 2016b). Many technical and scientific advances have accompanied this transition. This article draws the line between "old" and "new" models at the shift in priority from representing words to representing sentences. That is, while traditional DSMs seek a bottom-up story of semantics, which characterizes the lexicon and the composition functions that allow the lexical meanings to combine, newer methods favor a top-down approach, in which the goal is primarily to build systems capable of processing sentences, and lexical representations are viewed as emergent. This article refers to the relevant set of new models broadly as sentence encodings (SEs).

The top-down versus bottom-up distinction is not the only thing that differentiates modern SEs from traditional DSMs. Most notably, there has been a massive increase in model size; new SEs are often on the order of billions of parameters. As a result, it is debatable whether the observed successes of SEs are due to an improved ability to learn good linguistic representations or rather simply to their increased capacity to memorize co-occurrence patterns. Differentiating these two alternatives is challenging and often results in ambiguity about how exactly to interpret the empirical data on models' linguistic abilities (Section 4).

While a comprehensive review of SEs is impossible, several important variants, which are featured in the semantic analyses reviewed in Section 4, are summarized below.

2.2.1. Fixed-length sentence embeddings. The first SEs sought to directly extend word-embedding methods by treating a given sentence as a semantic unit, the meaning of which could

be defined by the context. A straightforward example is the SkipThought method (Kiros et al. 2015), which produces sentence embeddings using a neural network trained to predict a target sentence given a context sentence. Another notable method is InferSent (Conneau et al. 2017), in which representations are produced by a neural network optimized to perform natural language inference (NLI)—a task also known as recognizing textual entailment (Dagan et al. 2006), which requires predicting whether a premise p entails a hypothesis h.

Such embeddings have many advantages over the compositional DSMs described in Section 2.1. Since they are of fixed length, sentences of arbitrary syntactic structures can be easily compared with one another. And although such methods do not invoke explicit theories of syntactic composition, they do leverage order-aware neural networks (e.g., recurrent neural networks; Elman 1990), making them more syntax-aware than, for instance, naive vector addition.

2.2.2. Variable-length sentence representations. A massive paradigm shift occurred when fixed-length sentence embeddings were replaced by pretrained language models (LMs). LMs are artificial neural networks trained for the task of predicting the missing words in a sentence. For example, a basic LM is trained to predict the probability of the word w_t given all previous words $w_0, w_1, \ldots, w_{t-1}$. Pretraining refers to the process in which the parameters of a neural network are initialized by training on a task other than the eventual task of interest. In NLP, where tasks of interest include question answering, machine translation, and dialogue, language modeling has proven an effective pretraining task.

The first model to popularize pretrained LMs as a means of building SEs was Embeddings from Language Modeling (ELMo) (Peters et al. 2018a). ELMo is a two-layer recurrent neural network that, rather than learning a single vector to represent a sentence, produces two embeddings per word (one per layer) and considers the entire trained network to be a representation of the sentence. ELMo contrasts with earlier approaches in that it produces contextualized word representations—that is, word embeddings for each token rather than for each type. (Traditional embeddings as described in Section 2.1 are now often referred to as static embeddings to highlight this contrast.) ELMo's combination of LM pretraining and contextualized word embeddings resulted in unprecedented performance gains across a range of language processing tasks (Peters et al. 2018a).

The success of ELMo led to a flurry of research on similar pretrained LMs. Of these, the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al. 2018) has been the most extensively analyzed by those interested in syntactic and semantic representations. BERT differs from ELMo in two important ways. First, while ELMo uses a well-known recurrent neural network architecture, BERT uses a new architecture called the transformer (Vaswani et al. 2017). While recurrent models process their input in left-to-right order, transformers have no inherent inductive bias for linear order; instead, they process their input using a mechanism called "self-attention" (Vaswani et al. 2017), which effectively enables the model to induce arbitrary graph-like structures (not just before–after relations) over the input. Second, rather than standard language modeling, BERT uses two learning objectives: a masked language modeling (MLM) objective and a next sentence prediction (NSP) objective. MLM is essentially a cloze task in which the goal is to predict a word given its left and right context—for instance, *The [MASK] purred loudly* → *cat*. The NSP task aims to predict the next sentence given the current one.

Other noteworthy model variants are the Generative Pretrained Transformers (GPT2 and GPT3) (Radford et al. 2019, Brown et al. 2020). The GPT models, like ELMo, are trained on the task of left-to-right language modeling, but like BERT, they are based on the transformer architecture. The models have received attention largely due to their comparatively large size

(e.g., GPT3 has 175B parameters) and the impressive quality of text they are able to generate (Radford et al. 2020).

2.2.3. Training procedure and model size. There are several open questions about how training procedures influence the types of representations that models learn, and these issues potentially influence how we interpret the results reviewed in Section 4.

The first question concerns model size. Empirically, increased model size is associated with improved performance (Radford et al. 2019, Zhang et al. 2021). For example, Hu et al. (2020) and Warstadt et al. (2020) show that larger LMs better encode syntactic structures such as constituency and long-range agreements, and Peters et al. (2018a,b) and Tenney et al. (2019a) show that lower layers of networks appear to capture more syntactic information while higher layers capture more semantic information. Thus, the conventional wisdom is generally that models improve with size and depth and that results observed in small models are not necessarily indicative of how results would look if the model were scaled up. That said, the question is still wide open. We simply have not run the controlled experiments required to say whether large models' representations differ fundamentally from those of smaller models.

A second issue concerns the relationship between pretraining and fine-tuning—that is, the task-specific training that happens after pretraining when the model is optimized for a specific application, such as question answering. In general, fine-tuning after pretraining leads to better task performance than does training from scratch (i.e., training only on the target task without any pretraining) (Wang et al. 2019a). However, analysis also indicates that, after fine-tuning, the top layers of a network become specialized for the application and thus contain less linguistic information (Liu et al. 2019, Merchant et al. 2020, Mosbach et al. 2020). Such findings suggest that the conclusions (positive or negative) we draw from current experiments may differ from those we would draw if the same experiments were run using different fine-tuning regimens. Thus, again, it is best to consider current conclusions to be revisable: Future experiments may well provide new insights that force us to reinterpret the present data on syntactic and semantic representations.

3. EVALUATION METHODS

The question of how well deep learning models encode semantic structure presupposes a definition of semantic structure. Unfortunately, linguists, philosophers, and cognitive scientists have yet to reach a consensus. Thus, studies in NLP differ substantially in what they take to be evidence of semantics. Three broad categories of evaluation methods are summarized below.

3.1. Extrinsic Task-Based Evaluations

Since NLP's statistical revolution in the 1990s, the field has tended to prioritize standardized evaluations on benchmark tasks (Jurafsky & Martin 2009). This tradition extends to evaluations of semantic understanding in general. For example, Potts (2020) reframes the question of understanding as one of whether models can "achieve truly robust and general capabilities to answer questions, reason with language, and translate between languages." Few claim that performance on such tasks (question answering, NLI, machine translation) alone is sufficient to establish understanding, but many share the sentiment that if models are getting better at these challenging tasks, something about their semantic representations must be improving. As models have become more powerful, the trend has been away from single-task evaluations in favor of multitask evaluations (Nayak et al. 2016; Wang et al. 2019b,c), which require a single model to simultaneously perform well on a range of language tasks.

3.2. Targeted Task-Based Evaluations

In standard benchmarks, models are evaluated on samples that are drawn from the same distribution as the data on which the models are trained. As a result, such evaluations are likely to be gameable by models that memorize superficial statistical correlations, such as whether the word *sleeping* is more likely to occur in false sentences than in true ones (Poliak et al. 2018b). Such findings have turned many off of benchmark evaluations altogether (Linzen 2020, Bowman & Dahl 2021). Thus, increasingly, NLP research has turned to controlled ("targeted") evaluations, which are designed with the goal of testing whether models have encoded specific linguistic features (Linzen et al. 2016, Futrell et al. 2019, Linzen & Baroni 2020).

Most targeted semantic evaluations follow one of two formats. Those based on language modeling compare the probabilities that models assign to words given minimally differing contexts. For example, Ettinger (2020) evaluates models' understanding of lexical entailment by comparing the probabilities assigned to *bird* versus *tree* in the context *a robin is* {*not*} *a [MASK]*. This format allows pretrained LMs to be evaluated without any task-specific fine-tuning. However, not all SEs are LMs (see Section 2.2), and not all interesting semantic phenomena fit into the LM format. Thus, many semantic evaluations are instead based on NLI (Cooper et al. 1996, Poliak et al. 2018a, Rudinger et al. 2018, Kim et al. 2019). Such evaluations involve curated sets of premise/hypothesis pairs in which the inference hinges on the semantic phenomena of interest. For example, Dasgupta et al. (2018) use minimal pairs such as *the man is* {*not*} *taller than the woman/the woman is shorter than the man* to test models' understanding of negation.

3.3. Representational (Probing) Evaluations

Both extrinsic and targeted evaluations are cast in terms of model behavior—that is, what outputs do models produce for given inputs? Recently, there has been an increased interest in understanding and evaluating models directly in terms of the form of the model's internal representations (for a general review, see Belinkov & Glass 2019). Such evaluations seek evidence that the vector spaces underlying a model's inferences align to some predefined linguistic features. For example, the evaluations might ask whether the representations encode semantic roles, meaning that the model's representation of a given token (*man*) should differentiate whether it occurred as the agent (*The man saw the elephant*) or as the patient (*The man was seen by the elephant*) of a given verb (*to see*).

Among the most widely used techniques for inspecting representations is the probing classifier (Ettinger et al. 2016, Veldhoen et al. 2016, Adi et al. 2017, Conneau et al. 2018, Hupkes et al. 2018b), which attempts to predict a target linguistic feature given a representation. For example, given ELMo's vector representations of the words *man* and *to see* in the above sentences, the model would have to predict whether the word pair is an instance of the agent (arg0) semantic relation. If the classifier performs well on held-out words and sentences, then the representation is said to "encode" the notion of semantic agent. This method has been used to test whether fixed-length sentence embeddings encode information about word order and sentence length (Adi et al. 2017, Conneau et al. 2018) and whether contextualized word embeddings encode the range of syntactic and semantic features (Hewitt & Manning 2019; Tenney et al. 2019a,b). Note that, although widely used, probing classifiers are prone to false-positive results (Hewitt & Liang 2019, Voita & Titov 2020).

4. SEMANTIC STRUCTURE FROM DEEP LEARNING

This article focuses on three broad areas of semantic analysis: lexical semantics (Section 4.1), world knowledge (Section 4.2), and compositional semantics (Section 4.3). This is not an exhaustive

review but rather a distillation of the most salient evidence for and against claims that current deep learning representations perform well in each area. Notably absent is work on grounding and reference. While there exists work on multimodal variants of the models discussed here (Sun et al. 2019, Radford et al. 2021), such models have not undergone the same semantic analysis as text-only models. The lack of attention to grounding has attracted vocal criticism (Bender & Koller 2020), and future work in this direction should be a priority for computational semantics.

4.1. Lexicon and Lexical Semantics

Both traditional DSMs and early (fixed-length) SEs assume a fixed inventory of word representations that does not change from one sentence to the next. With the shift to contextualized word representations (ELMo, BERT), many have questioned whether newer models encode a recognizable lexicon at all—that is, a context-independent, type-level representation of the units from which sentences are composed. Attempts to answer this question have focused on two dimensions: the unit of representation within the lexicon (Section 4.1.1) and the syntactic-semantic structure of those units (Section 4.1.2). Work on lexical entailment and ontology is reviewed in Section 4.2.

4.1.1. Type representations. In this section, "type" refers to the token-type distinction for lexical items; for instance, in the sentence *The black cat chased the brown cat*, there are two tokens of a single *cat* type.

4.1.1.1. Evidence in favor. Vulić et al. (2020) are among the first to systematically explore how contextualized word embeddings (specifically, BERT) compare to static type-level DSMs [specifically, FastText (Bojanowski et al. 2017)] in terms of traditional lexical semantic evaluations. Such evaluations include word similarity (i.e., predicting a real-valued score for how "similar" two words are, which can be correlated with analogous human judgments), word analogies (e.g., filling in the blank given a pattern such as king:man::queen:__), and lexical relation prediction [e.g., differentiating synonymy versus hypernymy, where such relations are defined using resources such as WordNet (Fellbaum 2010)]. Vulić et al. (2020) estimate type representations from token representations by sampling word instances from the corpus at random and averaging their contextualized representations. Their results show that, in general, the best-performing contextualized models outperform the static embedding baselines, though there is high variance across languages. Parallel work by Chronis & Erk (2020) reports similar findings when using a different approach to estimate type representations, which directly accounts for words' multiple senses (e.g., the fact that *bank* may refer to either a riverside or a financial institution).

Looking specifically at word sense, a traditionally hard problem for static DSMs (Erk 2012), Peters et al. (2018a) and Reif et al. (2019) demonstrate that contextualized token embeddings encode sense naturally, performing competitively with prior state-of-the-art word sense disambiguation models despite no explicit training. Nair et al. (2020) further argue that the continuous notion of word sense captured by token embeddings is an improvement over that found in traditional lexical semantic resources; they use human studies to show that similarities generated by contextualized token embeddings have a high correlation with human judgments—for instance, treating homonyms (river *bank* versus financial *bank*) as less similar than polysemes (*chicken* animal versus *chicken* meat).

4.1.1.2. Evidence against. Mickus et al. (2020), like Chronis & Erk (2020) and Vulić et al. (2020), investigate whether BERT embeddings yield convincing type representations. Although they report similarly positive results on lexical similarity benchmarks, their overall conclusions

are far less optimistic. Using clustering analyses, the study finds that tokens of the same type do not cluster cohesively: 25% of tokens cluster as though they belonged to a different type than they do. While some of the effect can be explained by word sense, much is due to undesirable artifacts of training—for example, differentiating word tokens that appear in the first versus the second sentence of BERT's NSP objective (see Section 2.2.2). Yenicelik et al. (2020) find evidence of similar overdifferentiation within word senses. Using a combination of probing classifiers and visualization techniques, the study measures how well BERT embeddings align with WordNet's (Fellbaum 2010) sense inventory, showing that although accuracy is high, the space's organization is "not purely determined by semantics. . .[but is] intertwined with concepts such as syntax and sentiment" (Yenicelik et al. 2020, p. 160)—for instance, forming distinct clusters for the same sense of the word *arms*, one with positive sentiment (*swooped up into her arms. . .*) and one with negative (*. . .handcuffed arms. . .*). Giulianelli et al. (2020) observe similar results. However, both Yenicelik et al. (2020) and Giulianelli et al. (2020) are ultimately positive, arguing (as in Nair et al. 2020) that BERT's representations are an improvement over the hard-coded senses developed by linguists (see Section 4.1.3).

4.1.2. Argument structure. In this section, "argument structure" refers generally to models' ability to encode information about the semantic role of a noun (e.g., whether it is the argument of a verb, whether it is agent versus patient) and/or the syntactic-semantic constraints on a verb's use.

4.1.2.1. Evidence in favor. Ettinger (2020) uses psycholinguistic stimuli to test whether BERT recognizes anomalous reversals of semantic roles—for instance, differentiating minimally differing sentences such as *the restaurant owner forgot which customer/waitress the waitress/customer had served*. BERT performs well, ranking expected completions over unlikely ones 86% of the time (although it is worth noting that BERT's top-ranked prediction rarely matches the human's top-ranked choice). Studies that use probing classifiers also conclude that models like ELMo and BERT capture information about event and argument structure. Tenney et al. (2019a,b) show that such models capture semantic roles significantly better than a static word embeddings baseline. Li et al. (2021) show that entity representations are updated with state information as a consequence of the verbs with which they appear; for instance, the representation of *chest* in the context *you remove the key from the chest* contains information about whether `empty(chest)` is true.

Thrush et al. (2020) provide a thorough analysis of BERT's ability to represent semantic verb classes, focusing on selectional preferences and subcategorization frames. The authors use nonce words to control for confounds and then test the model's ability to generalize across an alternation pair. For example, during training, the model will see verbs from the same class (*spray/load*), some of which appear in only one alternation (e.g., *she will spray the wall with paint*). At test time, verbs appear in the unseen alternation (*she will spray paint onto the wall*). The study finds that, in aggregate, BERT assigns significantly higher probability to valid verb alternations than to invalid ones, suggesting that it generalizes syntactic behavior to words within the same class.

4.1.2.2. Evidence against. The strongest arguments that models do not encode argument structure rely on the fact that models do not seem to use such features during inference. McCoy et al. (2019), for example, show that BERT-based NLI models rely on naive heuristics such as whether sentences exhibit high lexical overlap, and Sinha et al. (2021) show that such models perform perfectly well even when words in a sentence are scrambled; both studies suggest that models ignore the argument structure that they purportedly encode.

Other arguments stem from the instability of results across experimental setups. For example, the above positive results are mostly based on pretrained LMs (in particular, BERT), but earlier studies of fixed-length sentence embeddings (Section 2.2.1) produce more negative findings. For example, Ettinger et al. (2016, 2018) use probing classifiers to evaluate whether fixed-length sentence embeddings can differentiate agents and patients of a given verb. They find that none of the models tested perform well: Accuracies are in the mid-60s compared with a 50% baseline. Even work on pretrained LMs is not consistently positive. Kann et al. (2019), like Thrush et al. (2020), investigate how well models capture verb alternations, but overall they are more negative in their interpretation, reporting at-chance performance for several verb classes. Their experimental setup uses a less standard sentence encoder and a different task design, making it difficult to compare with Thrush et al.'s (2020) positive findings.

4.1.3. Summary and discussion. With the shift to contextualized word representations (Section 2.2), many have asked, Do these modern SEs encode a lexicon at all? Taking the results discussed in Sections 4.1.1 and 4.1.2 together, this article views the answer as more positive than negative. Multiple independent studies confirm that contextualized representations outperform context-invariant ones on traditional metrics such as word similarity, analogies, and lexical entailment. There is also converging evidence from targeted evaluations and probing classifiers that contextualized word embeddings account for word sense and argument structure better than do traditional type-level DSMs.

Several important negative results must be accounted for. However, these should not necessarily be taken as evidence against the quality of SEs' lexical representations but rather as indicators of related theoretical and methodological gaps in need of further work.

The first major criticism is overdifferentiation: Type representations often map one-to-many or many-to-one onto the notions of word meanings defined by linguistic resources. While this clearly indicates that contextualized word embeddings induce a different lexicon than that encoded by existing resources, we cannot say yet whether this is a feature or a bug; that is, it remains to be determined whether contextualized embedding spaces provide a better or worse model of human lexical representations compared with static DSMs or explicit ontologies. Contextualized LMs generate predictions—for instance, about the roles that sentiment and discourse play in a token's representation—that could be formalized and validated in humans. Conducting such studies will require deeper collaboration between those analyzing SEs and those in theoretical and experimental semantics but will likely yield valuable insights. Frege himself argues that "it is enough if the sentence as whole has meaning; thereby also its parts obtain their meanings" (Frege 1884, quoted in Szabó 2020). Contextualized embeddings provide a concrete computational model within which to test such ideas.

The second major criticism is models' failures at inference time: Lexicosemantic structure (e.g., semantic roles) does not influence models' inferences in practice (e.g., in NLI) in the way it should. However, the relationship between the models' internal representations and their downstream behavior is not yet well understood within machine learning. Intuitively, training SEs involves training two subprocesses in tandem: (*a*) converting raw input (text) into representations and (*b*) converting representations into behavior (e.g., inferences). To a large extent, semantic analyses of SEs can treat these subprocesses as distinct. The positive results summarized above suggest that subprocess *a* is functioning comparably well. Given subprocess *a*, the observation that models often fail at inference time suggests a problem in subprocess *b*. Diagnosing and fixing problems in subprocess *b* is necessary if SEs are ever to be proposed as a complete model of semantic understanding. However, failing to have done so (yet) need not prevent us in the meantime

from using subprocess *a* to generate interesting insights about the structure of lexical semantic representations.

4.2. World Knowledge

Work in NLP often differentiates between linguistic knowledge and world knowledge. For example, one might falsely state that the capital of Michigan is Detroit, not because they do not understand lexical and compositional semantics but because they are simply misinformed about the world. Although the line between linguistic and world knowledge is difficult to draw rigorously (Hagoort & van Berkum 2007), NLP analyses are often framed in terms of this intuitive distinction. This section covers three categories of world knowledge: Section 4.2.1 covers encyclopedic knowledge (i.e., factoids such as that Obama was the 44th president of the United States) and commonsense knowledge (i.e., properties of common nouns, such as that robins are birds); Section 4.2.2 covers numeric reasoning.

4.2.1. Commonsense and encyclopedic knowledge. Although work focusing on world knowledge often further differentiates between encyclopedic and commonsense knowledge, the methods used to study both types of knowledge, and the conclusions drawn, have largely converged. Thus, this article discusses them together.

4.2.1.1. Evidence in favor. The seminal work investigating encyclopedic knowledge in SEs is by Petroni et al. (2019), who argue that pretrained LMs are on par with prior state-of-the-art knowledge bases at recalling facts about the world. The study uses prompts such as *Dante was born in [MASK]* to solicit answers from LMs and compares them with traditional knowledge bases of explicit tuples (e.g., ⟨Dante, born in, Italy⟩) extracted automatically from text (Etzioni et al. 2008). Petroni et al. (2019) do not claim that LMs are at human level—overall performance is in fact quite low, with the accuracy of the top-ranked prediction often below 20%—but rather that such performance is competitive with the best automatically constructed knowledge bases despite the fact that LMs have not been explicitly designed for this task.

On commonsense knowledge, Ettinger (2020) analyzes how well BERT encodes hyponym–hypernym relations using prompts such as *a robin is a [MASK]*. On prompts involving prototypical hypernyms, BERT performs well (e.g., ranking *bird* over alternatives like *tree* 100% of the time). Da & Kasai (2019) obtain similarly strong results using a different data set and probing design, but they still report a median of 100% accuracy when associating nouns with taxonomic relations (e.g., *is a fruit*, *is a toy*) compared with 89% for static (GLoVE) vectors (Pennington et al. 2014). Da & Kasai (2019) look beyond prototypical properties, using objects such as *zebra* and properties such as *is upright*. They use probing classifiers to measure whether contextualized embeddings of object names capture those objects' properties, and they find that BERT achieves median accuracies around 80%, nearly twice that of static embeddings.

Forbes et al. (2019) conduct a similar investigation using probing classifiers but focusing specifically on physical common sense about object attributes and affordances. The authors find that BERT performs consistently well at associating objects with attributes (e.g., that *apples* are *edible*) and attributes with affordances (e.g., that *edible* objects afford *eating*), often within 10 points (F measure) of human performance and in some cases exceeding it. BERT performs significantly less well at associating objects with affordances (e.g., that *apples* afford *eating*), though human performance on this task is also slightly lower. A related line of work uses pretrained LMs to augment human-interpretable resources like ConceptNet (Liu & Singh 2004), either by scoring (Li et al. 2016) or generating (Bosselut et al. 2019) novel commonsense tuples (e.g., that *mango* is a *fruit*).

Hwang et al. (2020) report that GPT3 produces such tuples with 73% accuracy, suggesting that pretrained LMs not only recognize that entities are related but can differentiate how they are related (e.g., *is a* versus *has a*).

4.2.1.2. Evidence against. One of the strongest pieces of evidence that SEs do not capture world knowledge is the fact that they can readily be made to generate false statements. Kassner & Schütze (2020) and Ettinger (2020), for example, show that models are insensitive to negation—for instance, readily producing *bird* when given the prompt *a robin is not a [MASK]*. Kassner & Schütze (2020) further show BERT to be overly sensitive to "misprimes." For example, a sentence such as *Talk? Birds can [MASK]* leads models to produce *talk* rather than the more acceptable response *fly*. Ravichander et al. (2020) report similar oversensitivity when changing from singular (*a robin is a [MASK]*) to plural (*robins are [MASK]*). Others simply point to embarrassing model failures as proofs-of-concept. For example, Ettinger (2020) prompts BERT with *The snow had piled up on the drive so high that they couldn't get the car out. When Albert woke up, his father handed him a [MASK]*, to which BERT offers the options *note*, *letter*, and *gun*.

Another primary criticism of prompt-based evaluation of LMs' knowledge is variability; for instance, a model might produce the correct answer when given *[MASK] is the capital of Rhode Island* but not when given *Rhode Island's capital is [MASK]*. Cao et al. (2021) conduct extensive experiments to show that models' predictions at test time closely track the distribution of answers in training regardless of test distribution and that predictions are by-and-large unchanged even when the relevant entities are masked out, suggesting that models' predictions are based more on the artifacts of the prompt than on the specific entities in question.

4.2.2. Numeric reasoning. A parallel line of work on world knowledge has asked whether LMs encode basic knowledge of numericity.

4.2.2.1. Evidence in favor. Johnson et al. (2020), studying a range of languages, find that models are generally able to differentiate grammatical numbers from ungrammatical ones (e.g., *two hundred three and fifty*). Wallace et al. (2019b) use probing classifiers trained for tasks such as predicting the largest number in a list to analyze how well embeddings of number tokens (e.g., *53*) capture information about magnitude and order. They report that static word embeddings often outperform BERT and in many cases perform close to an estimated upper bound. However, they note that models cannot extrapolate beyond the numeric range seen in training, a limitation well documented elsewhere (Saxton et al. 2019).

4.2.2.2. Evidence against. Subsequent studies, which use slightly different experimental designs, have failed to reproduce Wallace et al.'s (2019b) conclusions. For example, Johnson et al. (2020) use spelled-out words (*two* instead of *2*) and numbers greater than 100 and do not find that embeddings support value comparisons. Naik et al. (2019) consider yet another variation in design, which uses contrastive tuples of the form (*A*, *B*, *C*) such that a model must predict "true" if *A* is closer to *B* than *A* is to *C*. They find that embeddings are able to capture rough orders of magnitude (three versus one thousand) but not finer-grained distinctions (three versus four). Lin et al. (2020) show that BERT performs poorly on numeric MLM prompts (e.g., *a bird usually has [MASK] legs* → *two*).

4.2.3. Summary and discussion. In aggregate, the results on world knowledge are mixed. Results on numeric reasoning are more positive than perhaps would have been expected but are still negative overall. The more interesting discussion is about commonsense and encyclopedic

knowledge. On this topic, this article views the results as mostly positive, with a rationale echoing that in Section 4.1.3 on lexical semantics: Again, the strongest negative evidence presents a valid case against SEs as complete models of semantics, but it does not necessarily suggest that the conceptual structures themselves are incorrectly encoded by SEs.

The first major criticism is models' sensitivity to input: Models are just as likely to assert false facts as they are to assert truthful ones (e.g., ignoring negations and being distracted by misprimes). While models indeed make these errors, studies of semantics have long assumed a separation between the system that assigns meanings to words and the system that composes those meanings in order to process sentences. Before SEs, representing world knowledge entailed the use of explicit datastores (Liu & Singh 2004, Carlson et al. 2010, Fellbaum 2010), which were evaluated based on whether they connected entities via the correct semantically typed relations (Berant et al. 2011, Hosseini et al. 2018). While SEs clearly fail to process conceptual relations compositionally, there is little evidence to suggest that the SEs are failing to correctly encode the conceptual relations themselves. Rather, SEs appear not only to relate entities but also to differentiate between types of relations (e.g., *is-a* versus *has-a*). Debates about how much we can isolate conceptual knowledge from compositional semantics are worthwhile, but such debates concern the semantic processing pipeline in general, not the quality of SEs specifically.

As in Section 4.1.3, a second criticism concerns failures at inference time: There are just too many examples of SEs generating nonsensical narratives. Again, such criticisms get at the underlying question of how deep learning models convert internal conceptual representations into behavior. The importance of this question should not be minimized. Our inability to answer this is at the heart of why current SEs are simply language processing systems rather than models of semantics. However, when trying to understand which aspects of semantic structure SEs potentially account for, it is important to differentiate the representations models learn from the tasks they are optimized to perform. Ettinger (2020, p. 46) summarizes the problem well: "Whereas the function of language processing for humans is to compute meaning and make judgments of truth, language models. . .simply leverage the most reliable cues in order to optimize their predictive capacity." It is in principle possible that SEs could encode correct conceptual relations and nonetheless produce nonsense output, if the most salient features of the nonsense better matched sentences found in corpora than did the most salient features of true sentences. Thus, focusing primarily on behavior at inference time arguably amounts to conflating performance with competence and can distract from models' progress at encoding conceptual structure under the hood.

4.3. Compositionality

Among the most central tenets of formal semantics is the principle of compositionality: that the meaning of a sentence should be a function of the meaning of the words and the way in which they are combined (Partee 1995). Unlike traditional DSMs, in which composition functions are defined explicitly, SEs learn to combine words implicitly while performing tasks like language modeling. Thus, it is debatable whether SEs have the type of composition functions often assumed in formal linguistics. There is no standard test for compositional behavior, and existing analysis work makes use of a variety of definitions. This article summarizes work focusing on three criteria that have received particular attention: systematicity (Section 4.3.1), negation (Section 4.3.2), and phrase representations (Section 4.3.3).

4.3.1. Systematicity. In Fodor & Pylyshyn's (1988, p. 37) words, systematicity is the property that guarantees that "the ability to produce/understand some utterances is intrinsically

connected to the ability to produce/understand certain others"; for instance, a system capable of understanding the sentence *John loves Mary* should by definition be capable of understanding *Mary loves John*. While it is a matter of debate how systematic human language processing is, it is generally agreed that language requires some degree of systematic generalization and that SEs should exhibit the capacity for such.

4.3.1.1. Evidence against. Most sections in this article lead with positive results. However, as the consensus is that SEs do not behave systematically, it makes sense to lead here with the negative results.

In one of the first studies to investigate systematicity, Lake & Baroni (2018) define a toy task in which models receive natural language instructions as input (*turn left then jump twice*) and are required to produce corresponding sequences of symbolic actions as output ("LTURN JUMP JUMP"). They train a recurrent model from scratch to perform this task, and they evaluate it on a test set defined such that good performance cannot be achieved by pure memorization (e.g., a model might be trained on {*jump twice*, *walk then turn left*} and then tested on {*turn left twice*}). The study finds that test sets that require generalizing to significantly longer sequences than have been seen in training lead models to "fail spectacularly" (Lake & Baroni 2018, p. 2880). Kim & Linzen (2020) report similar findings when using a simple semantic parsing task—for instance, $a\ cat\ smiled \rightarrow cat(x1) \wedge smile(x2) \wedge agent(x1, x2)$—showing (again) that models trained from scratch fail to extrapolate to longer sentences/deeper trees and that they generalize poorly when known words appear in previously unseen roles (e.g., *cat* appearing as a patient having previously been observed only as an agent). Goodwin et al. (2020) draw similarly negative conclusions in yet a third study design, which uses jabberwocky sentences to test whether SEs learn context-invariant inference patterns that generalize to arbitrary content words ($all\ blickets\ dax \rightarrow \neg some\ blickets\ don't\ dax$). They find comparably low performance and high variance when tested on sentence pairs that require reasoning about function words in the context of novel content words, suggesting that the models do not generalize as desired. Finally, while the three preceding studies use toy task settings and examine the behavior of models trained from scratch on the task of interest, Yanaka et al. (2019) document a similar lack of systematicity in pretrained LMs (BERT), using an NLI test set that requires systematic application of monotonicity rules.

A separate set of evidence that pretrained LMs behave unsystematically comes not from linguistically motivated research but rather from related research on robustness and trustworthiness in AI. Such work shows that even state-of-the-art models often change their predictions as a result of seemingly meaningless changes, such as replacing a word with a synonym ($movie \rightarrow film$) (Ribeiro et al. 2018, 2020) or appending nonsense sequences of words to the input (Wallace et al. 2019a).

4.3.1.2. Evidence in favor. Given the above, it is hard to argue that SEs are systematic learners in the way symbolic models of logical reasoning are. However, it is important not to oversimplify the picture. While all of the authors above are ultimately negative about models' generalization abilities, it is only in some experiments that models "fail spectacularly." In many other cases, the results are open to more charitable interpretations. For example, Goodwin et al.'s (2020) primary conclusion is based on the observation that performance on inferences involving unseen content words is lower than performance on inferences involving seen words. However, performance on the generalization test set is only slightly below that on the in-distribution set (accuracy in the 80s versus the 90s), meaning that models still behave correctly more often than not. Results are also stronger when models are trained with more data and when both sentences contain

identical content words (as in the *all blickets dax*→ ¬*some blickets don't dax* example), suggesting that the negative results could be explained by a failure to learn lexical categories rather than by a failure to apply rules systematically. Similarly, neither Lake & Baroni's (2018) nor Kim & Linzen's (2020) results are uniformly negative. Lake & Baroni (2018) report that when novel inputs can be understood using mix-and-match strategies, simple recurrent models are able to perform quite well. When Kim & Linzen (2020) break down performance by model and type of generalization, they find that transformers in particular (as opposed to recurrent models) perform well on active–passive transformations and decently on subject–object translations (e.g., *John loves Mary* → *Mary loves John*), though admittedly with high variation across configurations.

Finally, some authors are overtly positive about the potential of such models to behave systematically, especially if one considers hybrid neurosymbolic variants. For example, Hupkes et al. (2018a), like Lake & Baroni (2018), study the performance of recurrent neural models using a toy compositionality task. The study finds that simple changes to training (specifically, introducing supervision on the attention layer) can significantly improve the models' generalization performance. Conklin et al. (2021) show that meta-learning (i.e., optimizing training for out-of-distribution generalization) yields significant improvements on the tasks from the studies of both Lake & Baroni (2018) and Kim & Linzen (2020).

4.3.2. Negation. Negation is arguably the most frequently cited phenomenon in discussions of models' need to process language compositionally.

4.3.2.1. Evidence in favor. Early work on negation in SEs (Ettinger et al. 2018) uses probing classifiers to test whether models trained on semantic tasks (e.g., NLI) yield representations distinguishing verbs that appear within the scope of a negation from verbs that appear in a negated sentence but outside of the negation scope (e.g., *professor* versus *student* in *the professor is not helping the student*). Ettinger et al. (2018) find that most of the SEs tested perform quite well (accuracy in the high 90s), though they acknowledge that the test set is not free of confounds.

Other studies have investigated models' sensitivity to negation via their treatment of negative polarity items (NPIs)—that is, lexical items like *any*, which are grammatical only when used in downward entailing contexts, such as under the scope of negation. Jumelet & Hupkes (2018) find that a simple recurrent model trained from scratch is consistently less surprised by sentences containing grammatical NPI uses than by those containing ungrammatical uses. Warstadt et al. (2019) perform a similar study and find that BERT performs close to perfectly. Slight variations on the task, however, weaken the results significantly—for instance, if BERT is asked to make an absolute binary prediction (grammatical/ungrammatical) rather than relative ranking (more/less grammatical).

4.3.2.2. Evidence against. The task of negation scope detection, on which all of the above evaluations rely, has been argued to be more syntactic than semantic (McKenna & Steedman 2020). Thus, many who are interested in negation focus instead on the question of whether models' inferences are influenced by negation in the correct way. One of the earliest such studies, by Dasgupta et al. (2018), focuses on the InferSent model (Section 2.2.1). The study uses a targeted NLI test set consisting of simple sentence pairs (*the woman is more cheerful than the man/the man is not more cheerful than the woman*) that are balanced such that simple word-level heuristics would perform at chance. Off the shelf, InferSent performs poorly but better than random (50% accuracy over a 33% random baseline). The authors interpret this result as evidence that the model can exploit signal beyond naive unigrams but that it falls short of real compositional reasoning.

Work on pretrained LMs has shown a similar lack of sensitivity to negation. For example, Kassner & Schütze (2020) and Ettinger (2020) find that while LMs recall basic facts with high precision (*birds can fly*), they are equally likely to generate the complete opposite "fact" (*birds cannot fly*), seemingly ignoring the presence of the negation entirely.

4.3.3. Phrase representations. Another line of work has explored compositionality by looking at representations of short phrases, in a manner reminiscent of earlier work on DSMs (Section 2.1)—for instance, *coffee cake*, *hit ball*.

4.3.3.1. Evidence in favor. Shwartz & Dagan (2019) perform a comprehensive analysis to compare contextualized word embeddings with static ones in terms of their ability to represent short phrases. The study uses six different lexical composition tasks (e.g., detecting verb-particle constructions, detecting nonliteral noun compounds) and finds that, on average, contextualized representations perform better at "detecting meaning shift," in particular on tasks involving light verb constructions (e.g., differentiating *carry it* from *carry on*) (Shwartz & Dagan 2019, p. 403).

4.3.3.2. Evidence against. Most studies yield negative results about SEs' ability to model phrases. Early analyses, predating pretrained LMs, report that networks trained on NLI from scratch fail to reason about adjective–noun compounds (e.g., *little baby*) (Pavlick & Callison-Burch 2016) and simple lexical alternations (e.g., *playing a guitar* versus *playing a flute*) (Glockner et al. 2018). More recently, Yu & Ettinger (2020, 2021) have used BERT and similar models to score the similarity of phrase pairs (e.g., *average person/ordinary citizen*); they find that, after controlling for naive word overlap, correlations with human similarity judgments are very low, suggesting "little evidence of nuanced composition" (Yu & Ettinger 2020, p. 4896). Nandakumar et al. (2019) evaluate multiple SEs on the task of scoring the degree of compositionality for a given noun phrase (e.g., *application form* versus *sitting duck*). Across all evaluations, ELMo and BERT embedding models perform worse than a static word-embedding baseline that treats phrases as atomic units (e.g., *sitting_duck*). Garcia et al. (2021) draw similar conclusions. Finally, Shwartz & Dagan (2019) find that contextualized word embeddings struggle to infer implicit meaning in phrases—for instance, failing to correctly paraphrase noun compounds such as *olive oil* → *oil made from olives*.

4.3.4. Summary and discussion. In general, the consensus is that models are not compositional (or at least not compositional enough). Thus, while modern SEs appear to be good at modeling word meaning (Sections 4.1.3 and 4.2.3), they struggle with composition. Such a one-line summary could have been made of the DSMs of 10 years ago, but it would be wrong to say that no progress has been made. The question of whether models behave compositionally has come to the fore and has sustained interest because, despite the negative results above, SEs simply do not fail spectacularly enough to close the book on the issue. Perhaps the strongest positive evidence is the existence proof of models performing well on so many classically hard language tasks (question answering, NLI, machine translation), on which traditional DSMs were never even viable competitors. Even granting issues with extrinsic evaluations (Section 3), it is hard to deny that SEs are doing something that is new and exciting.

Thus, the most salient question is less about behavior (what SEs do) and more about evaluation (what they should do). Without a precise enough definition of what type of composition would be human-level, we cannot currently make claims, positive or negative, about SEs as models of compositional semantics. Generating converging evidence in either direction requires agreeing on priorities along at least two dimensions.

4.3.4.1. Controlled or realistic? The above studies of compositionality differ significantly in ecological validity. The strongest negative results rely on idealized data in which there is a fully transparent relationship between syntax and semantics. Such studies offer good insights about how neural networks learn, but they are limited in what they can tell us about SEs as models of semantics in general for two reasons.

First, whether human semantic inferences behave systematically is an open question rather than something to be taken for granted. Recent work provides numerous examples of humans making seemingly unsystematic judgments about entailment on both natural (De Marneffe et al. 2019, Ross & Pavlick 2019) and highly controlled (White & Rawlins 2018) stimuli. As such, results on idealized data frequently fail to predict how well SEs will explain real human judgments. In a study by Ettinger (2020), models perform poorly on recognizing negation when given templatic sentences (*A robin is not a tree*) but much better when given sentences based on real corpora (*most smokers find that quitting isn't very effective*). In a study by Yanaka et al. (2019), models produce different error patterns depending on whether they are tested on realistic monotonicity inferences (generated via crowdsourcing) or templatic ones (sourced from linguistics textbooks).

Second, evaluations on simulated languages typically train models from scratch on the target task. However, general (task-nonspecific) language exposure is an important part of human language development, and empirical evidence suggests that pretraining endows models with inductive biases above and beyond what is provided by the model architecture alone (Warstadt et al. 2020, Lovering et al. 2021).

Thus, there is reason to be skeptical that the negative results observed in from-scratch models in simulation are representative of how modern SEs behave on natural language. At best, such studies tell us that SEs are not implementations of particular existing theories of compositional semantics. They do not help us assess SEs' potential as alternative models of compositional semantics. Future work needs to consider both possibilities (Section 5).

4.3.4.2. Representation or behavior? A larger, more philosophical question prevents us from drawing stronger conclusions from the presented studies. In particular, we lack consensus on whether compositionality should be understood primarily as a property of a model's representations or of its behavior. Work in logic and philosophy frequently defines compositionality in terms of a model's representation: It is a property of how semantic representations are structured, such that the structure of some concepts guarantees something about the structure of others (Fodor & Pylyshyn 1988). Such definitions say little about how models behave, and supporting arguments rely minimally on experimental data about human behavior. In contrast, work in experimental and computational linguistics often defines compositionality using human behavior rather than any explicit definition. Such work solicits humans' intuitive judgments about how words combine (Lapata & Lascarides 2003, De Marneffe et al. 2019) and seeks to provide explicit models that produce the same judgments.

Evaluative work on SEs has mixed and matched these approaches. For example, much of the work discussed in Section 4.3.1 cites Fodor & Pylyshyn (1988) but provides experimental results in terms of model behavior. Work discussed in Section 4.3.3, inspired by human behavioral data, places comparatively more emphasis on model representations (e.g., probing the embeddings themselves). Such cross-cutting evaluations are potentially informative if done deliberately, but in most current studies, it is difficult to determine which assumptions about compositionality underlie the interpretation of results.

Future work should bring our evaluations closer in line with the theories we seek to test. Such work will no doubt require proposing new ways of operationalizing existing theories and new ways of formalizing SEs' behavior, but this is worthwhile. Precisely connecting evaluation criteria

to theories of semantics will enable us to better understand the implications of SEs for the broader study of semantics.

5. CONCLUSION

Modern deep learning methods yield representations of words and sentences that encode significantly more conceptual information about lexical semantics (Section 4.1) and the external world (Section 4.2) than did earlier DSMs. However, they still fall short when reasoning about these concepts compositionally (Section 4.3). The same general summary—good at lexical semantics, bad at composition—could have been made of the DSMs of 10 years ago. However, it is an oversimplification to say that models have not changed. To the contrary, new models invoke fundamentally new notions of conceptual knowledge—for instance, eschewing explicit type-level lexicons (Section 4.1.3) and external knowledge bases (Section 4.2.3). Even the way in which new models struggle with composition is different from how traditional DSMs did (Section 4.3.4). Traditional compositional DSMs, inspired by formal semantics, showed promise in toy settings but failed to compete on realistic language-understanding tasks. Modern SEs, in contrast, show unprecedented, even human-level, performance on complex tasks such as reading comprehension and machine translation but break when tested in controlled simulations.

New SEs thus deserve to be considered with a fresh perspective. However, linguistic analysis has only just begun and has yet to yield theoretical insight. If SEs are to transition from being mere language processing systems to a genuine model of semantics, there are several overarching barriers that future work must overcome.

Perhaps the most important open question is, What should a computational model of semantics deliver? There are two obvious perspectives one could take. We can focus on the implementation level and ask, Can SEs be an implementation of a given, prespecified model of semantics? Alternatively, we can focus on the computational level and ask, Can we formalize the behavior of SEs into a new, alternative model of semantics? The former is obligated to specify a clear theory of semantics within which models are being assessed. The latter is obligated to generate explicit theories with testable hypotheses that can be validated in human subjects. Both approaches are valid and valuable; researchers need not commit themselves to one at the exclusion of the other, but individual studies should. Currently, studies are cast implicitly as one or the other—for instance, the studies described in Section 4.3.1 exemplify the former while the studies described in Section 4.1.1 exemplify the latter—but are not overt enough in such commitments to yield the type of rigorous theoretical contributions needed to propel the field forward.

A related question is, How do we prioritize model representations versus model behavior? Ultimately, we seek a complete theory of semantics that explains behavior in terms of representations, and representations in terms of learning. However, we are currently faced with complex deep learning models within which the precise interfaces between such stages are poorly understood. As a result, conclusions often differ based on the evaluation methods used—for instance, task performance versus probing analyses. This problem is not new—it echoes the long-known challenge of differentiating performance from competence. However, debates about SEs' semantic abilities are rarely framed as such. Again, casting studies in more explicit theoretical terms, making clear claims about theories of semantic representations versus models of human behavior, would make it easier to recognize converging evidence where it exists and to diagnose which aspects of a proposed model are weakest.

All considered, the opportunities to explore important new theories of semantic representations at the word and sentence levels are significant. Semanticists, theoretical as well as computational, have every reason to be excited and should welcome this new technology as an opportunity for deep, cross-disciplinary collaboration.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

Thank you to Sam Bowman, Yejin Choi, Katrin Erk, Allyson Ettinger, Tal Linzen, and Martha Palmer for invaluable feedback on earlier drafts. Thank you to the students of the Brown LUNAR Lab for work and discussions that informed the interpretations presented here. Factual errors and misinterpretations are entirely my own.

LITERATURE CITED

Adi Y, Kermany E, Belinkov Y, Lavi O, Goldberg Y. 2017. *Fine-grained analysis of sentence embeddings using auxiliary prediction tasks*. Paper presented at the International Conference on Learning Representations (ICLR 2017), Toulon, Fr., Apr. 24–26

Adiwardana D, Luong MT, So DR, Hall J, Fiedel N, et al. 2020. Towards a human-like open-domain chatbot. arXiv:2001.09977 [cs.CL]

Baroni M, Dinu G, Kruszewski G. 2014. Don't count, predict! A systematic comparison of context-counting versus context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Vol. 1: *Long Papers*, pp. 238–47. Stroudsburg, PA: Assoc. Comput. Linguist.

Baroni M, Zamparelli R. 2010. Nouns are vectors, adjectives are matrices: representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1183–93. Stroudsburg, PA: Assoc. Comput. Linguist.

Belinkov Y, Glass J. 2019. Analysis methods in neural language processing: a survey. *Trans. Assoc. Comput. Linguist.* 7:49–72

Bender EM, Koller A. 2020. Climbing towards NLU: on meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5185–98. Stroudsburg, PA: Assoc. Comput. Linguist.

Bengio Y, Ducharme R, Vincent P, Janvin C. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.* 3:1137–55

Berant J, Dagan I, Goldberger J. 2011. Global learning of typed entailment rules. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 610–19. Stroudsburg, PA: Assoc. Comput. Linguist.

Bojanowski P, Grave E, Joulin A, Mikolov T. 2017. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* 5:135–46

Boleda G. 2020. Distributional semantics and linguistic theory. *Annu. Rev. Linguist.* 6:213–34

Boleda G, Herbelot A. 2016. Formal distributional semantics: introduction to the special issue. *Comput. Linguist.* 42(4):619–35

Bosselut A, Rashkin H, Sap M, Malaviya C, Celikyilmaz A, Choi Y. 2019. COMET: commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4762–79. Stroudsburg, PA: Assoc. Comput. Linguist.

Bowman SR, Dahl GE. 2021. What will it take to fix benchmarking in natural language understanding? arXiv:2104.02145 [cs.CL]

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, et al. 2020. Language models are few-shot learners. arXiv:2005.14165 [cs.CL]

Cao B, Lin H, Han X, Sun L, Yan L, et al. 2021. Knowledgeable or educated guess? Revisiting language models as knowledge bases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Vol. 1: *Long Papers*, pp. 1860–74. Stroudsburg, PA: Assoc. Comput. Linguist.

Carlson A, Betteridge J, Kisiel B, Settles B, Hruschka E, Mitchell T. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, pp. 1306–13. Palo Alto, CA: AAAI Press

Chronis G, Erk K. 2020. When is a bishop not like a rook? When it's like a rabbi! Multi-prototype BERT embeddings for estimating semantic relationships. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pp. 227–44. Stroudsburg, PA: Assoc. Comput. Linguist.

Conklin H, Wang B, Smith K, Titov I. 2021. Meta-learning to compositionally generalize. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Vol. 1: *Long Papers*, pp. 3322–35. Stroudsburg, PA: Assoc. Comput. Linguist.

Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 670–80. Stroudsburg, PA: Assoc. Comput. Linguist.

Conneau A, Kruszewski G, Lample G, Barrault L, Baroni M. 2018. What you can cram into a single $&!#* vector: probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Vol. 1: *Long Papers*, pp. 2126–36. Stroudsburg, PA: Assoc. Comput. Linguist.

Cooper R, Crouch D, Van Eijck J, Fox C, Van Genabith J, et al. 1996. *Using the framework*. Tech. Rep. LRE 62-051 D-16, FraCaS Consort.

Da J, Kasai J. 2019. Cracking the contextual commonsense code: understanding commonsense reasoning aptitude of deep contextual representations. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pp. 1–12. Stroudsburg, PA: Assoc. Comput. Linguist.

Dagan I, Glickman O, Magnini B. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges: Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, pp. 177–90. Berlin: Springer

Dasgupta I, Guo D, Stuhlmüller A, Gershman SJ, Goodman ND. 2018. Evaluating compositionality in sentence embeddings. arXiv:1802.04302 [cs.CL]

De Marneffe MC, Simons M, Tonhauser J. 2019. The CommitmentBank: investigating projection in naturally occurring discourse. In *Proceedings of Sinn und Bedeutung 23*, pp. 107–24. Bellaterra, Spain: Univ. Autòn Barcelona

Devlin J, Chang MW, Lee K, Toutanova K. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 [cs.CL]

Dhillon PS, Foster D, Ungar L. 2011. Multi-view learning of word embeddings via CCA. *In NIPS'11: Proceedings of the 24th International Conference on Neural Information Processing Systems*, pp. 199–207. New York: Assoc. Comput. Mach.

Dumais ST, Furnas GW, Landauer TK, Deerwester S, Harshman R. 1988. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 281–85. New York: Assoc. Comput. Mach.

Elman JL. 1990. Finding structure in time. *Cogn. Sci.* 14(2):179–211

Emerson G. 2020. What are the goals of distributional semantics? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7436–53. Stroudsburg, PA: Assoc. Comput. Linguist.

Erk K. 2007. A simple, similarity-based model for selectional preferences. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 216–23. Stroudsburg, PA: Assoc. Comput. Linguist.

Erk K. 2012. Vector space models of word meaning and phrase meaning: a survey. *Lang. Linguist. Compass* 6(10):635–53

Ettinger A. 2020. What BERT is not: lessons from a new suite of psycholinguistic diagnostics for language models. *Trans. Assoc. Comput. Linguist.* 8:34–48

Ettinger A, Elgohary A, Phillips C, Resnik P. 2018. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1790–801. Stroudsburg, PA: Assoc. Comput. Linguist.

Ettinger A, Elgohary A, Resnik P. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 134–39. Stroudsburg, PA: Assoc. Comput. Linguist.

Etzioni O, Banko M, Soderland S, Weld DS. 2008. Open information extraction from the web. *Commun. ACM* 51(12):68–74

Fellbaum C. 2010. WordNet. In *Theory and Applications of Ontology: Computer Applications*, pp. 231–43. Dordrecht, Neth.: Springer

Firth JR. 1957. A synopsis of linguistic theory, 1930–1955. In *Studies in Linguistic Analysis*, pp. 1–32. Oxford, UK: Philol. Soc.

Fodor JA, Pylyshyn ZW. 1988. Connectionism and cognitive architecture: a critical analysis. *Cognition* 28(1–2):3–71

Forbes M, Holtzman A, Choi Y. 2019. Do neural language representations learn physical commonsense? arXiv:1908.02899 [cs.CL]

Frege G. 1884. *Die Grundlagen der Arithmetik: Eine logisch mathematische Untersuchung über den Begriff der Zahl.* Breslau: W. Koebner

Futrell R, Wilcox E, Morita T, Qian P, Ballesteros M, Levy R. 2019. Neural language models as psycholinguistic subjects: representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1: *Long and Short Papers*, pp. 32–42. Stroudsburg, PA: Assoc. Comput. Linguist.

Garcia M, Kramer Vieira T, Scarton C, Idiart M, Villavicencio A. 2021. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Vol. 1: *Long Papers*, pp. 2730–41. Stroudsburg, PA: Assoc. Comput. Linguist.

Giulianelli M, Del Tredici M, Fernández R. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3960–73. Stroudsburg, PA: Assoc. Comput. Linguist.

Glockner M, Shwartz V, Goldberg Y. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Vol. 2: *Short Papers*, pp. 650–55. Stroudsburg, PA: Assoc. Comput. Linguist.

Goodwin E, Sinha K, O'Donnell TJ. 2020. Probing linguistic systematicity. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1958–69. Stroudsburg, PA: Assoc. Comput. Linguist.

Grefenstette E, Sadrzadeh M. 2011. Experimenting with transitive verbs in a DisCoCat. arXiv:1107.3119 [cs.CL]

Hagoort P, van Berkum J. 2007. Beyond the sentence given. *Philos. Trans. R. Soc. B* 362(1481):801–11

Harris ZS. 1954. Distributional structure. *Word* 10(2–3):146–62

Hartung M, Kaupmann F, Jebbara S, Cimiano P. 2017. Learning compositionality functions on word embeddings for modelling attribute meaning in adjective-noun phrases. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Vol. 1: *Long Papers*, pp. 54–64. Stroudsburg, PA: Assoc. Comput. Linguist.

Hewitt J, Liang P. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–43. Stroudsburg, PA: Assoc. Comput. Linguist.

Hewitt J, Manning CD. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1: *Long and Short Papers*, pp. 4129–38. Stroudsburg, PA: Assoc. Comput. Linguist.

Hosseini MJ, Chambers N, Reddy S, Holt XR, Cohen SB, et al. 2018. Learning typed entailment graphs with global soft constraints. *Trans. Assoc. Comput. Linguist.* 6:703–17

Hu J, Gauthier J, Qian P, Wilcox E, Levy R. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1725–44. Stroudsburg, PA: Assoc. Comput. Linguist.

Hupkes D, Singh A, Korrel K, Kruszewski G, Bruni E. 2018a. Learning compositionally through attentive guidance. arXiv:1805.09657 [cs.CL]

Hupkes D, Veldhoen S, Zuidema W. 2018b. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *J. Artif. Intel. Res.* 61:907–26

Hwang JD, Bhagavatula C, Bras RL, Da J, Sakaguchi K, et al. 2020. COMET-ATOMIC 2020: on symbolic and neural commonsense knowledge graphs. arXiv:2010.05953 [cs.CL]

Johnson D, Mak D, Barker A, Loessberg-Zahl L. 2020. Probing for multilingual numerical understanding in transformer-based language models. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 184–92. Stroudsburg, PA: Assoc. Comput. Linguist.

Jumelet J, Hupkes D. 2018. Do language models understand anything? On the ability of LSTMs to understand negative polarity items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 222–31. Stroudsburg, PA: Assoc. Comput. Linguist.

Jurafsky D, Martin JH. 2009. *Speech and Language Processing*. New York: Prentice-Hall, Inc. 2nd ed.

Kann K, Warstadt A, Williams A, Bowman SR. 2019. Verb argument structure alternations in word and sentence embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pp. 287–97. Stroudsburg, PA: Assoc. Comput. Linguist.

Kassner N, Schütze H. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7811–18. Stroudsburg, PA: Assoc. Comput. Linguist.

Kim N, Linzen T. 2020. COGS: a compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9087–105. Stroudsburg, PA: Assoc. Comput. Linguist.

Kim N, Patel R, Poliak A, Xia P, Wang A, et al. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pp. 235–49. Stroudsburg, PA: Assoc. Comput. Linguist.

Kiros R, Zhu Y, Salakhutdinov R, Zemel RS, Torralba A, et al. 2015. Skip-thought vectors. arXiv:1506.06726 [cs.CL]

Lake B, Baroni M. 2018. Generalization without systematicity: on the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, ed. J Dy, A Krause, pp. 2873–82. n.p.: PMLR

Landauer TK, Dumais ST. 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* 104(2):211–40

Lapata M, Lascarides A. 2003. A probabilistic account of logical metonymy. *Comput. Linguist.* 29(2):261–315

LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521(7553):436–44

Lenci A. 2008. Distributional semantics in linguistic and cognitive research. *Ital. J. Linguist.* 20(1):1–31

Lenci A. 2018. Distributional models of word meaning. *Annu. Rev. Linguist.* 4:151–71

Li BZ, Nye M, Andreas J. 2021. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Vol. 1: *Long Papers*, pp. 1813–27. Stroudsburg, PA: Assoc. Comput. Linguist.

Li X, Taheri A, Tu L, Gimpel K. 2016. Commonsense knowledge base completion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Vol. 1: *Long Papers*, pp. 1445–55. Stroudsburg, PA: Assoc. Comput. Linguist.

Lin BY, Lee S, Khanna R, Ren X. 2020. Birds have four legs?! NumerSense: probing numerical commonsense knowledge of pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6862–68. Stroudsburg, PA: Assoc. Comput. Linguist.

Linzen T. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5210–17. Stroudsburg, PA: Assoc. Comput. Linguist.

Linzen T, Baroni M. 2020. Syntactic structure from deep learning. *Annu. Rev. Linguist.* 7:195–212

Linzen T, Dupoux E, Goldberg Y. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Trans. Assoc. Comput. Linguist.* 4:521–35

Liu H, Singh P. 2004. ConceptNet—a practical commonsense reasoning tool-kit. *BT Technol. J.* 22(4):211–26

Liu NF, Gardner M, Belinkov Y, Peters ME, Smith NA. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1: *Long and Short Papers*, pp. 1073–94. Stroudsburg, PA: Assoc. Comput. Linguist.

Lovering C, Jha R, Linzen T, Pavlick E. 2021. *Predicting inductive biases of pre-trained models*. Poster presented at the International Conference on Learning Representations (ICLR 2021), Vienna, Austria, May 4

Lund K, Burgess C, Atchley R. 1995. Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, ed. JD Moore, JF Lehman, pp. 660–65. Mahwah, NJ: Erlbaum

McCoy T, Pavlick E, Linzen T. 2019. Right for the wrong reasons: diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–48. Stroudsburg, PA: Assoc. Comput. Linguist.

McKenna N, Steedman M. 2020. Learning negation scope from syntactic structure. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pp. 137–42. Stroudsburg, PA: Assoc. Comput. Linguist.

Merchant A, Rahimtoroghi E, Pavlick E, Tenney I. 2020. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 33–44. Stroudsburg, PA: Assoc. Comput. Linguist.

Mickus T, Paperno D, Constant M, van Deemter K. 2020. What do you mean, BERT? In *Proceedings of the Society for Computation in Linguistics 2020*, pp. 279–90. Stroudsburg, PA: Assoc. Comput. Linguist.

Mikolov T, Chen K, Corrado G, Dean J. 2013. Efficient estimation of word representations in vector space. arXiv:1301.3781 [cs.CL]

Mitchell J, Lapata M. 2010. Composition in distributional models of semantics. *Cogn. Sci.* 34(8):1388–429

Mosbach M, Khokhlova A, Hedderich MA, Klakow D. 2020. On the interplay between fine-tuning and sentence-level probing for linguistic knowledge in pre-trained transformers. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 68–82. Stroudsburg, PA: Assoc. Comput. Linguist.

Naik A, Ravichander A, Rose C, Hovy E. 2019. Exploring numeracy in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3374–80. Stroudsburg, PA: Assoc. Comput. Linguist.

Nair S, Srinivasan M, Meylan S. 2020. Contextualized word embeddings encode aspects of human-like word sense knowledge. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pp. 129–41. Stroudsburg, PA: Assoc. Comput. Linguist.

Nandakumar N, Baldwin T, Salehi B. 2019. How well do embedding models capture non-compositionality? A view from multiword expressions. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pp. 27–34. Stroudsburg, PA: Assoc. Comput. Linguist.

Nayak N, Angeli G, Manning CD. 2016. Evaluating word embeddings using a representative suite of practical tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 19–23. Stroudsburg, PA: Assoc. Comput. Linguist.

Partee B. 1995. Lexical semantics and compositionality. In *Invitation to Cognitive Science*, Vol. 1: *Language*, ed. L Gleitman, M Liberman, DN Osherson, pp. 311–60. Cambridge, MA: MIT Press. 2nd ed.

Pavlick E, Callison-Burch C. 2016. Most "babies" are "little" and most "problems" are "huge": compositional entailment in adjective-nouns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Vol. 1: *Long Papers*, pp. 2164–73. Stroudsburg, PA: Assoc. Comput. Linguist.

Pennington J, Socher R, Manning CD. 2014. GloVe: global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–43. Stroudsburg, PA: Assoc. Comput. Linguist.

Peters M, Neumann M, Iyyer M, Gardner M, Clark C, et al. 2018a. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1: *Long Papers*, pp. 2227–37. Stroudsburg, PA: Assoc. Comput. Linguist.

Peters M, Neumann M, Zettlemoyer L, Yih W-T. 2018b. Dissecting contextual word embeddings: architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1499–509. Stroudsburg, PA: Assoc. Comput. Linguist.

Petroni F, Rocktäschel T, Riedel S, Lewis P, Bakhtin A, et al. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–73. Stroudsburg, PA: Assoc. Comput. Linguist.

Poliak A, Haldar A, Rudinger R, Hu JE, Pavlick E, et al. 2018a. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 67–81. Stroudsburg, PA: Assoc. Comput. Linguist.

Poliak A, Naradowsky J, Haldar A, Rudinger R, Van Durme B. 2018b. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 180–91. Stroudsburg, PA: Assoc. Comput. Linguist.

Potts C. 2020. Is it possible for language models to achieve language understanding? *Medium*, Oct. 5. **https://chrisgpotts.medium.com/is-it-possible-for-language-models-to-achieve-language-understanding-81df45082ee2**

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, et al. 2021. Learning transferable visual models from natural language supervision. arXiv:2103.00020 [cs.CV]

Radford A, Wu J, Amodei D, Amodei D, Clark J, et al. 2020. Better language models and their implications. *OpenAI Blog*, Feb. 14. **https://openai.com/blog/better-language-models/**

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. 2019. *Language models are unsupervised multitask learners*. Work. Pap., OpenAI, San Francisco

Rajpurkar P, Zhang J, Lopyrev K, Liang P. 2016a. SQuAD: 100,000+ questions for machine comprehension of text. arXiv:1606.05250 [cs.CL]

Rajpurkar P, Zhang J, Lopyrev K, Liang P. 2016b. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2383–92. Stroudsburg, PA: Assoc. Comput. Linguist.

Ravichander A, Hovy E, Suleman K, Trischler A, Cheung JCK. 2020. On the systematicity of probing contextualized word representations: the case of hypernymy in BERT. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pp. 88–102. Stroudsburg, PA: Assoc. Comput. Linguist.

Reif E, Yuan A, Wattenberg M, Viegas FB, Coenen A, et al. 2019. Visualizing and measuring the geometry of BERT. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, ed. H Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, E Fox, R Garnett. Red Hook, NY: Curran Assoc. **https://papers.nips.cc/paper/2019/file/159c1ffe5b61b41b3c4d8f4c2150f6c4-Paper.pdf**

Ribeiro MT, Singh S, Guestrin C. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistic*s, Vol. 1: *Long Papers*, pp. 856–65. Stroudsburg, PA: Assoc. Comput. Linguist.

Ribeiro MT, Wu T, Guestrin C, Singh S. 2020. Beyond accuracy: behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–12. Stroudsburg, PA: Assoc. Comput. Linguist.

Ross A, Pavlick E. 2019. How well do NLI models capture verb veridicality? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2230–40. Stroudsburg, PA: Assoc. Comput. Linguist.

Rudinger R, Naradowsky J, Leonard B, Van Durme B. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 2: *Short Papers*, pp. 8–14. Stroudsburg, PA: Assoc. Comput. Linguist.

Saxton D, Grefenstette E, Hill F, Kohli P. 2019. *Analysing mathematical reasoning abilities of neural models*. Paper presented at the International Conference on Learning Representations (ICLR 2019), New Orleans, LA, May 6–9

Shwartz V, Dagan I. 2019. Still a pain in the neck: evaluating text representations on lexical composition. *Trans. Assoc. Comput. Linguist.* 7:403–19

Sinha K, Parthasarathi P, Pineau J, Williams A. 2021. UnNatural Language Inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Vol. 1: *Long Papers*, pp. 7329–46. Stroudsburg, PA: Assoc. Comput. Linguist.

Smolensky P. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artif. Intel.* 46(1–2):159–216

Sun C, Myers A, Vondrick C, Murphy K, Schmid C. 2019. VideoBERT: a joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7464–73. Los Alamitos, CA: IEEE

Szabó ZG. 2020. Compositionality. In *Stanford Encyclopedia of Philosophy*, ed. EN Zalta. Stanford, CA: Stanford Univ. **https://plato.stanford.edu/entries/compositionality/**

Tenney I, Das D, Pavlick E. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–601. Stroudsburg, PA: Assoc. Comput. Linguist.

Tenney I, Xia P, Chen B, Wang A, Poliak A, et al. 2019b. *What do you learn from context? Probing for sentence structure in contextualized word representations*. Paper presented at the International Conference on Learning Representations (ICLR 2019), New Orleans, LA, May 6–9

Thrush T, Wilcox E, Levy R. 2020. Investigating novel verb learning in BERT: selectional preference classes and alternation-based syntactic generalization. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 265–75. Stroudsburg, PA: Assoc. Comput. Linguist.

Turney PD, Pantel P. 2010. From frequency to meaning: vector space models of semantics. *J. Artif. Intel. Res.* 37:141–88

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. 2017. Attention is all you need. arXiv:1706.03762 [cs.CL]

Veldhoen S, Hupkes D, Zuidema W. 2016. Diagnostic classifiers: revealing how neural networks process hierarchical structure. In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches (CoCo 2016)*, ed. TR Besold, A Bordes, A d'Avila Garcez, G Wayne. Aachen, Ger.: RWTH Aachen Univ. **http://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper6.pdf**

Voita E, Titov I. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–96. Stroudsburg, PA: Assoc. Comput. Linguist.

Vulić I, Ponti EM, Litschko R, Glavaš G, Korhonen A. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7222–40. Stroudsburg, PA: Assoc. Comput. Linguist.

Wallace E, Feng S, Kandpal N, Gardner M, Singh S. 2019a. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2153–62. Stroudsburg, PA: Assoc. Comput. Linguist.

Wallace E, Wang Y, Li S, Singh S, Gardner M. 2019b. Do NLP models know numbers? Probing numeracy in embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5307–15. Stroudsburg, PA: Assoc. Comput. Linguist.

Wang A, Hula J, Xia P, Pappagari R, McCoy RT, et al. 2019a. Can you tell me how to get past Sesame Street? Sentence-level pretraining beyond language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4465–76. Stroudsburg, PA: Assoc. Comput. Linguist.

Wang A, Pruksachatkun Y, Nangia N, Singh A, Michael J, et al. 2019b. SuperGLUE: a stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, ed. H Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, E Fox, R Garnett. Red Hook, NY: Curran Assoc. **https://papers.nips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf**

Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR. 2019c. *GLUE: a multi-task benchmark and analysis platform for natural language understanding*. Paper presented at the International Conference on Learning Representations (ICLR 2019), New Orleans, LA

Warstadt A, Cao Y, Grosu I, Peng W, Blix H, et al. 2019. Investigating BERT's knowledge of language: five analysis methods with NPIs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2877–87. Stroudsburg, PA: Assoc. Comput. Linguist.

Warstadt A, Zhang Y, Li X, Liu H, Bowman SR. 2020. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 217–35. Stroudsburg, PA: Assoc. Comput. Linguist.

White AS, Rawlins K. 2018. The role of veridicality and factivity in clause selection. In *Proceedings of the 48th Annual Meeting of the North East Linguistic Society (NELS 48)*, pp. 221–34. Amherst, MA: Grad. Linguist. Stud. Assoc.

Yanaka H, Mineshima K, Bekki D, Inui K, Sekine S, et al. 2019. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 31–40. Stroudsburg, PA: Assoc. Comput. Linguist.

Yenicelik D, Schmidt F, Kilcher Y. 2020. How does BERT capture semantics? A closer look at polysemous words. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 156–62. Stroudsburg, PA: Assoc. Comput. Linguist.

Yu L, Ettinger A. 2020. Assessing phrasal representation and composition in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4896–907. Stroudsburg, PA: Assoc. Comput. Linguist.

Yu L, Ettinger A. 2021. On the interplay between fine-tuning and composition in transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2279–93. Stroudsburg, PA: Assoc. Comput. Linguist.

Zhang Y, Warstadt A, Li X, Bowman SR. 2021. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Vol. 1: *Long Papers*, pp. 1112–25. Stroudsburg, PA: Assoc. Comput. Linguist.