Evaluating Compositionality in Sentence Embeddings

Ishita Dasgupta[1], **Demi Guo**[2], **Andreas Stuhlmüller**[3], **Samuel J. Gershman**[4] & **Noah D. Goodman**[3]

[1]Department of Physics and Center for Brain Science, Harvard University
[2]Department of Computer Science, Harvard University
[3]Departments of Psychology and Computer Science, Stanford University
[4]Department of Psychology and Center for Brain Science, Harvard University

Abstract

An important challenge for human-like AI is compositional semantics. Recent research has attempted to address this by using deep neural networks to learn vector space embeddings of sentences, which then serve as input to other tasks. We present a new dataset for one such task, "natural language inference" (NLI), that cannot be solved using only word-level knowledge and requires some compositionality. We find that the performance of state of the art sentence embeddings (InferSent; Conneau et al., 2017) on our new dataset is poor. We analyze the decision rules learned by InferSent and find that they are largely driven by simple heuristics that are ecologically valid in its training dataset. Further, we find that augmenting training with our dataset improves test performance on our dataset without loss of performance on the original training dataset. This highlights the importance of structured datasets in better understanding and improving AI systems.

Keywords: Sentence embeddings; compositionality; test datasets

Introduction

A hallmark of human intelligence is compositionality: the ability, in the words of von Humboldt, to "make infinite use of finite means." The failure of neural network models to achieve compositionality has been a recurring (and controversial) theme in cognitive science (Fodor & Pylyshyn, 1988; Gershman & Tenenbaum, 2015; Lake & Baroni, 2017). However, recent successes of powerful deep learning systems trained on large corpora have renewed hopes that neural networks can close the gap with humans. In this paper, we explore minimal cases in a "natural language inference" (NLI, MacCartney (2009); Dagan et al. (2006)) task that cannot be solved without taking compositional information into account and thus develop a stringent test for compositionality. We then ask to what extent the state-of-the-art system for performing this task exhibits a truly compositional understanding of natural language.

Our approach is motivated partly by the need for better benchmarks to assess AI systems (White et al., 2017; Marelli et al., 2014; Pavlick & Callison-Burch, 2016; Gershman & Tenenbaum, 2015). Currently, most systems are trained and evaluated on large corpora which can be partially gamed by simple heuristics. For example, Socher et al. (2011) presented a recursive autoencoder that achieved state-of-the-art performance on paraphrase detection, yet it only performed 10% better than a baseline method that simply reported the most frequent class. The fact that these highly sophisticated algorithms may only be doing slightly better than naive baselines is brought into focus by more diagnostic benchmarks. We see a role for cognitive science in designing benchmarks that better probe the competences of AI systems, much in the same way that cognitive scientists have been probing the competences of humans (Ritter et al., 2017; Lake et al., 2018).

Our results show that while the system we test exhibits poor performance on our compositional test set, much of its failure can be traced to biases in the training dataset. Furthermore, we see that the system is capable of exhibiting some compositionality given the right training data, pointing to potential uses for such structured datasets not just as diagnostic tools, but also for improving training of models.

Background

Sentence Embeddings

Vector-based models of word semantics have been successful in capturing many aspects of word meanings. However, understanding language requires not only understanding words, but understanding their relations within a sentence. Due to the combinatorial productivity of language, the number of possible sentences far exceeds the size of the vocabulary; therefore generating similar vector embeddings for sentences has proven challenging. Recent literature reports several supervised as well as unsupervised approaches to learning sentence representations using Recurrent neural networks (RNNs) that account for word ordering (Kiros et al., 2015; Hill et al., 2016; Conneau et al., 2017). These are intended to capture semantic content, and do perform reasonably well on transfer tasks—i.e. other sentence-level tasks which the embeddings were not specifically trained on. Particularly, the performance of these sentence models exceeds the performance of bag-of-words models that patently lack any relational information about the words (i.e., compositionality). However, it is unclear exactly what compositional information is gained in RNN sentence models beyond lexical meaning.

Natural Language Inference classifiers

The sentence embeddings we explore in this paper are from InferSent (Conneau et al., 2017). We choose to use these sentence embeddings as they represent the current state-of-the-art for transfer in semantic tasks, and we expect that strong performance in transfer tasks indicates a good representation for the semantics of a sentence. These embeddings were trained end-to-end using the architecture in Figure 1 on the SNLI (Stanford Natural Language Inference) training set (Bowman et al., 2015). The training task is to classify pairs of sentences into 'entailment', 'contradiction', or 'neutral'.

The embeddings were shown to perform well on other tasks (such as sentiment analysis, semantic textual similarity and other natural language inference datasets) by re-using the embedding layers and training only the classifier for the specific task at hand. We train the model using the same protocol as in Conneau et al. (2017) for use in this work. Our trained InferSent model gives us 84.73% accuracy on validation and 84.84% accuracy on the SNLI test set, which is comparable the performance of the classifier reported in Conneau et al. (2017). For comparison, we also train a bag of words (BOW) baseline model that averages the GloVe embeddings (Pennington et al., 2014) for all the words in the sentence to form a sentence embedding. We train a multi-layer perceptron on these embeddings to give the BOW-MLP classifier we use in the following. BOW-MLP achieves 53.99% accuracy on the SNLI test set (comparable to the BOW performance reported in Conneau et al., 2017).
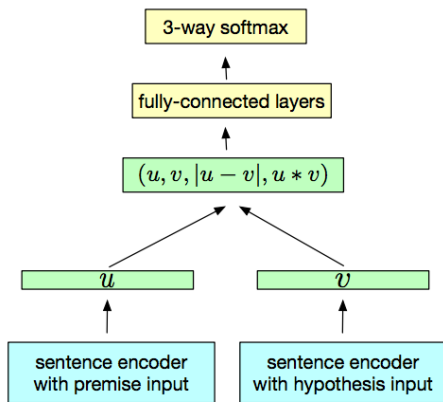

Figure 1: InferSent architecture (Conneau et al., 2017).

SNLI dataset

The Stanford Natural Language Inference dataset (Bowman et al., 2015) is a large annotated corpus for NLI that is generated with a crowdsourcing framework. Workers are presented with a scene description from a corpus of image captions, and asked to supply sentences that have each of three possible relations (entailment, neutral, and contradiction) to the given sentence. The freedom to produce entirely novel sentences leads to a rich set of examples from the set of possible sentences; however, it also leads to some unexpected biases that we will discuss in later sections.

Notion of Compositionality

Compositionality can mean many things. The notion that we focus on for this work is the abstract understanding of how words combine, in a way that generalizes to words and phrases that have not previously been encountered. For example, rules of the type in Table 1 hold true for X, Y and Z that may never have been encountered in that combination before. In fact, it should generalize to X, Y and Z that have never been encountered before at all. Understanding this sort

of abstract rule, for any combinatorially large possible values for X, Y and Z, is a step to a more general understanding of compositional representations of sentence structure.

Compositional comparisons dataset

Our goal is to design pairs of sentences such that the NLI relation within a pair (entailment, neutral or contradiction) can be changed without changing the words involved, simply by changing the word ordering within each sentence. We thus generate sets of sentence pairs which differ by permutation of words, such that the pairs represent different relations.

Type	Entailment hypothesis	Contradiction hypothesis	# of pairs
Same	X is more Y than Z	Z is more Y than X	14670
More-Less	Z is less Y than X	X is less Y than Z	14670
Not	Z is not more Y than X	X is not more Y than Z	14670

Table 1: Comparisons dataset summary. Set of rules for premise: X is more Y than Z

By construction, BOW models will perform at chance on this task, since they cannot distinguish the pairs. This provides a hard baseline for the performance that is possible without abstract rule understanding. In the literature, any performance above a BOW model is often seen as proof of compositionality. However, this is an unwarranted conclusion—the BOW model baseline usually receives only averaged word vectors for the sentence and therefore theoretically also loses some of the lexical information. We propose to instead gauge the compositionality of sentence-vector models by seeing how differently they classify these permuted sets.

We generate our test dataset using comparisons, as these yield many simple examples of sentence pairs that require more than word-level data to understand (when comparing two entities, their order in the sentence matters), and generation of several such sentence pairs can be easily automated. We consider three sub-types, described below and summarized in Table 1.

Same type

A-B pairs differ only in the order of the words.
```
A: The woman is more cheerful than the man
B: The man is more cheerful than the woman
CONTRADICTION
A: The woman is more cheerful than the man
B: The woman is more cheerful than the man
ENTAILMENT
```

More-Less type

A-B pairs differ by whether they contain the word 'more' or the word 'less'.
```
A: The woman is more cheerful than the man
B: The woman is less cheerful than the man
CONTRADICTION
A: The woman is more cheerful than the man
B: The man is less cheerful than the woman
ENTAILMENT
```

Not type

A-B pairs differ by whether they contain the word 'not'.

```
A: The woman is more cheerful than the man
B: The woman is not more cheerful than the man
CONTRADICTION
A: The woman is more cheerful than the man
B: The man is not more cheerful than the woman
ENTAILMENT
```

To facilitate comparison with the SNLI dataset, we ensure that the vocabulary distribution of the Comparisons dataset is similar to the original SNLI training dataset. Only a few words differ by more than 1% from their occurrence rate in SNLI, such as *not, a, than, the, is, less, more*. This is inevitable given the general structure of the comparison sentence pairs we use.

Classification Analysis

The overall performance of each of the classifiers on the Comparisons dataset are given in Table 2.

Type	BOW-MLP	InferSent
same	50.0	50.37
more/less	30.24	50.35
not	48.98	45.24

Table 2: Performance on the Comparisons dataset.

BOW-MLP

As expected, BOW-MLP makes classifications that are exactly symmetric across the two true categories in each task, since members of each category are just permuted versions of each other and BOW cannot distinguish them (Figure 2). This also ensures that the performance is capped at 50%. A sign of using more than word-level information would be asymmetry between the classifications of the two categories.

InferSent

The performance of InferSent is slightly more asymmetric (Figure 3), indicating that it is able to use some information beyond the word level. Yet overall InferSent is extremely poor at this task, indicating that it fails to fully exhibit the compositionality needed for these comparison sentences. We next analyze some of the patterns of classification errors observed.

All same words When the words in both sentences are the same (the same-type comparisons) they are largely classified as entailments (Figure 3), despite half being true contradictions. We observe that in the SNLI dataset, most contradictory sentence pairs have no overlap in words. For example, a contradictory sentence pair in SNLI is:

```
A: Several people are trying to climb a ladder
in a tree.
B: People are watching a ball game.
```
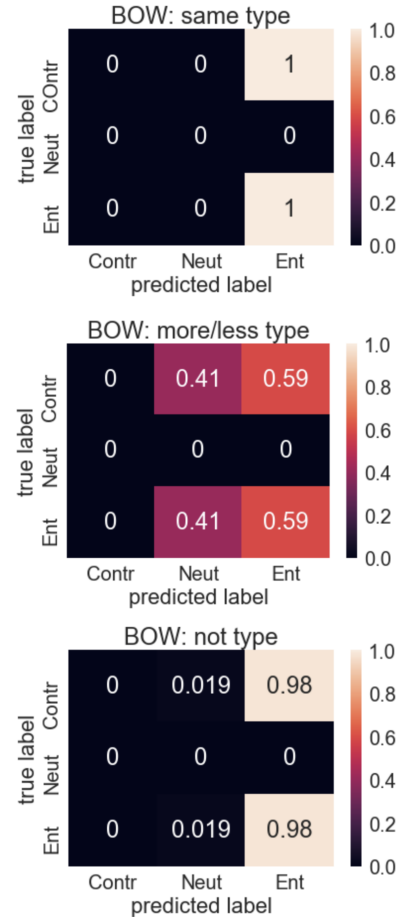

Figure 2: BOW-MLP confusion matrices, with normalized rows.

```
CONTRADICTION
```

Thus, within SNLI, it is much more likely for a sentence pair to be entailment or neutral if they have significant overlap. In order to quantitatively verify this observation, we rank all the sentence pairs by overlap rate: $\frac{\text{\# of overlap words}}{\text{total \# of words}}$ (in non-increasing order). We then look at the top X sentences with highest overlap. As shown in Table 3, 91.5% of the pairs with top 1000 maximum overlap between the sentences have the true label of either entailment or neutral, and are very rarely true contradiction.

Top	Entailment	Neutral	Contradiction
All	33.4%	33.3%	33.3%
10000	39.5%	35.7%	24.8%
1000	50.8%	40.7%	8.5%

Table 3: Overlap rate of words in SNLI.

Thus, InferSent may be learning the heuristic that high overlap in words predicts entailment, rather than a compositional semantic representation. This explains the failure of
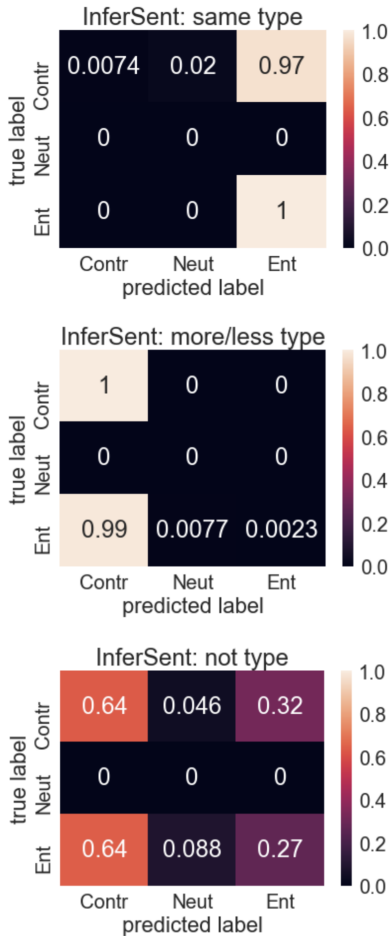
Figure 3: InferSent confusion matrices, with normalized rows.

InferSent to generalize to our same-type task.

Difference of one word When the words in two sentences differ by just one word, the decision is largely based on whether those words have opposing meanings irrespective of the order of the words. We see this from performance on more-less type comparisons (Figure 3). Here the words across the pairs differ only in the presence or absence of the word 'more' or 'less'. Since the relation between the words 'more' and 'less' is largely contradictory, we hypothesize that their use in a pair of sentences leads the classifier to presume the sentences are contradictory, irrespective of the order of the words.

We evaluate this hypothesis by investigating the statistics of antonyms in the SNLI dataset. To check whether a sentence pair (A,B) contains antonyms, we go through each word in sentence A, and consider all synonyms of that word, and consider all antonyms of those synonyms. Finally, we check if sentence B contains any of those antonyms.

We observe that this heuristic is fairly consistent with the SNLI data. Table 4 shows that the presence of antonyms strongly predicts a true label of contradiction in the SNLI dataset (61.2% compared to chance at 33.3%). We also see

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

Table 4: Antonym word pair in the SNLI dataset

that a true contradiction predicts the presence of an antonym pair (12.2%) more strongly than entailment does (3.5%).

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

Table 5: Antonym word pair in high overlap subset of SNLI.

Since most of our dataset contains a large amount of overlap between sentences A and B, we check for statistics of the high overlap set as well (Table 5). Here, the trend of a true contradiction predicting the presence of an antonym pair (43.7%) more strongly than entailment does (8.7%), is more pronounced. However, the presence of an antonym pair no longer predicts the true label of contradiction at a high rate. This is likely due to the very low base rate of contradictions in the high overlap subset of SNLI, as compared to entailments.

Overall, these results suggest, again, that the underlying statistics of the SNLI dataset lead models, including InferSent, to ignore the order of words when solving our more/less-type task.

Negations Comparisons that differ in the presence or absence of the negation 'not' are preferentially classified as contradictions (Figure 3). To verify that this heuristic is largely consistent with the SNLI dataset, we look at sentence pairs that contain "negating N-grams": no, not, n't. (By considering "n't", we will consider words such as "don't" or "doesn't".)

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

Table 6: Negation in SNLI dataset.

We observe that a "negation difference yields contradiction" heuristic is consistent with the SNLI data. Table 6 shows that the presence of a negation strongly predicts a true label of contradiction in the SNLI dataset (58.4% compared to chance at 33.3%). We also see that a true contradiction predicts the presence of an antonym pair (3.3%) slightly more strongly than entailment does (1.1%).

We repeat the analysis for the top 10,000 of the high overlap set as well (Table 7). Here, the presence of negation predicts a contradiction even more strongly than in the full dataset (despite the lower base rates of contradiction in this subset of the data), indicating a very strong basis for this heuristic within the high overlap subset of the SNLI dataset.

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

Table 7: Negation in the high overlap subset of SNLI.

Summary of heuristics We find evidence for a few heuristics that explain the bulk of the patterns seen in the performance of InferSent on our Comparisons dataset, all of which have ecological validity in the SNLI dataset. First, we find that a large overlap in words between two sentences leads InferSent to believe that they entail one another. Second, we see that the difference of one word between the two sentences, when the difference is an antonym or a negation, leads InferSent to classify them as contradictions irrespective of word order. Both of these illustrate a disproportionate dependence on lexical, rather than compositional meaning in InferSent.

The analysis so far has highlighted word-level heuristics that InferSent might be using. Yet the confusion matrix results (Figure 3) show a slight asymmetry, indicating at least minor multi-word effects. One hypothesis is that larger deviations in the order of overlapping words, alone, leads InferSent to dissfavor entailment. This is trivially true for same-type comparisons where the exact same word order results in an entailment inference, and different word order sometimes leads to other classifications (top row of the same-type comparisons in Figure 3). But in this case these are the correct classifications, so the heuristic is indistinguishable from full compositional reasoning. Critically, in the case of comparatives of the 'not' type, pairs that differ more in the word order are in fact entailments. We see that for this type of example, InferSent classifies true contradictions as entailments *more* than it does true entailments ($p = 0.2e - 11$).

This suggests, though certainly doesn't prove, a heuristic that differing word order in the presence of 'order-promoting' words like 'more' and 'less' like in our Comparisons dataset, disfavors entailment judgments. There are other simple uses of word order that could be in play: for instance, antomynic pairs of bigrams could generalize the single-word heuristics described above. However, a systematic analysis of the effect of word order, and of the ecological validity of such heuristics, is challenging due to the combinatorial explosion in the number of possibilities. We leave a thorough investigation of this to future work.

Augmented training

The foregoing results suggest that biases in the SNLI training data may be enough to lead a sentence encoding model to use simple non-compositional representations. This leaves open the question of whether architectures such as InferSent are capable of representing the relational features needed to succeed at our Compositional task. In this section, we explore this question by retraining the InferSent model on a combined dataset which includes both the Comparisons dataset and original SNLI training data. This serves to test whether

simple training on examples of the rules in Table 1, will enable InferSent to generalize these rules to X, Y and Z that it has previously never seen in that combination. This is a step towards gauging the compositionality of this sentence representation.

The training subset of our Comparisons dataset consists of 40k sentence pairs (7% of the 550k pair SNLI training set). Validation and test sets consist of 2000 sentence pairs each. There is no overlap between any of these sets.

Fine-tuning

We first tried initializing with the model trained on the SNLI dataset, and then training it on our new Comparisons dataset (using the same protocols used in Conneau et al. (2017) to train InferSent). Results are shown in Table 8. We observed that model performance on the SNLI data task degrades over the course of training (test accuracy went from 84.84 % to 56.37 %), though it remained higher than the random baseline of 33.3 %. The final model, however, performs very well on the Comparisons dataset (99.8 % test accuracy).

So while a decline in the performance of the model on SNLI points to over-fitting to the data, we see that the model doesn't simply memorize the specific training data used from the Comparisons dataset, and does actually learn the compositional rules (as evidenced by high test and validation performance on the Comparisons dataset). This indicates that the InferSent model architecture is in theory able to learn such relational patterns, given the right training data.

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

Table 8: Experiment: Finetuning

Retraining

To explore whether it is possible to perform well on both the Comparisons and SNLI datasets, we next trained a model from scratch on a training dataset that includes both SNLI and our Comparisons dataset, again using the same training protocol used in the original paper on InferSent (Conneau et al., 2017). Results are shown in Table 9. The test accuracy on SNLI (84.96 %) is comparable to the model trained only on SNLI (84.84 %). Moreover, test accuracy on the Comparisons dataset (99.55 %) is much higher than the model trained only on SNLI (45.36 %). Thus we show that it is possible to train a model such that it has high performance on specially designed edge-cases like the Comparisons dataset, without loss of performance on the more general SNLI dataset.

Epoch	Train(Combined)	Test(SNLI)	Test(Comp)
12	90.99%	84.96%	100.00%

Table 9: Experiment: Retraining From Scratch

This result also verifies our previous hypothesis that the model learns the simplest ways to accommodate the training data: the main reason the InferSent model performs badly on the Comparisons dataset is that its training data licenses "shortcut" biases, not because of shortcomings in the model itself. This points to the benefits of understanding potential biases in training data and including specially designed data to correct them.

Discussion

This work highlights the inadequacy of mainstream tasks in truly testing if Natural Language Processing (NLP) models represent compositional structure beyond the word level. InferSent achieves high performance on the test set of the SNLI dataset, as well as several other transfer tasks, but fails on our Comparisons dataset. This indicates that it misses crucial aspects of the compositionality in sentence meaning. How then does the InferSent model succeed on SNLI? Analysis of the behavior of the model on our well-controlled dataset allowed us to conjecture some word-driven heuristics, many of which we found have ecological validity in the SNLI training data.

This points to the utility of carefully designed datasets both for testing models' representational abilities, as well as for better understanding what they have actually learned. This is especially useful for models with large parameter spaces and many local minima, where heuristic solutions can explain much of the variance in the training data.

Elucidating the blind spots in a system's encoding of compositionality can then be utilized to improve it. We found that the InferSent model can be trained to perform better on our Comparisons comparisons without reducing performance on SNLI, by just including a part of the comparison dataset in the training data. This indicates that, for this case, the shortcoming is not in the model architecture, but rather in the poverty and biases of the training data. By debiasing training corpora and augmenting them with minimal contrasting examples, we can move closer to a truly compositional encoding of language.

Future Directions

Our Comparisons dataset has the crucial property that, by construction, it cannot be solved with only word-level information. Building a more general Comparisons dataset with this property that extends beyond comparison-type sentences is an important direction for future research. Another clear direction is to assess how other models, such as SkipThought (Kiros et al., 2015), perform on these problems, and explore the heuristics they might be covertly employing. Using techniques for generating interpretable explanations from classification decisions (e.g. Ribeiro et al., 2016) could help to better understand the strengths and weaknesses of these models on diagnostic datasets; and in turn perhaps prescribe new training regimes.

Further work on augmented training will be needed to better isolate the benefits of including specially designed data in training: do the results transfer to other tasks that require similar aspects of compositionality or even to more distant aspects of understanding beyond the word level?

References

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised Learning of Universal Sentence Representations from Natural Language Inference Data.

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 tectual entailment*, (pp. 177–190). Springer.

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

Gershman, S., & Tenenbaum, J. B. (2015). Phrase similarity in humans and machines. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

Hill, F., Cho, K., & Korhonen, A. (2016). Learning Distributed Representations of Sentences from Unlabelled Data.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors. In *Advances in neural information processing systems*, (pp. 3294–3302).

Lake, B. M., & Baroni, M. (2017). Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks. *arXiv preprint arXiv:1711.00350*.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2018). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40.

MacCartney, B. (2009). *Natural language inference*. Stanford University.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., Zamparelli, R., et al. (2014). A sick cure for the evaluation of compositional distributional semantic models. In *LREC*, (pp. 216–223).

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 (Volume 1: Long Papers)*, vol. 1, (pp. 2164–2173).

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, (pp. 1532–1543).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 1135–1144). ACM.

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

Socher, R., Huang, E. H., Pennin, J., Manning, C. D., & Ng, A. Y. (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in Neural Information Processing Systems*, (pp. 801–809).

White, A. S., Rastogi, P., Duh, K., & Van Durme, B. (2017). Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, vol. 1, (pp. 996–1005).