Gauging Compositionality in Sentence Embeddings

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

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

Abstract

Word embeddings trained on large corpora are frequently used as features for NLP systems. Recent research has attempted to generate such vector space embeddings for larger pieces of text, such as sentences and paragraphs. The understanding of language and the combinatorially large number of possible sentences hinges on the ability to understand the systematic compositionality of how words combine. While these vector representations of sentences outperform bag-of-words models on several tasks, it is unclear how much compositionality they truly capture. We present a dataset for a natural language inference task that cannot be solved using only lexical/word level knowledge and instead requires varying degrees of compositionality. We use sentence embeddings from InferSent (Conneau et al., 2017), the state of the art in sentence level transfer tasks, and find that performance on our tasks is poor, indicating that the representations induced by this model capture little compositionality. We analyze what limited compositionality is in fact learned by InferSent (affording it higher performance on NLI than bag-of-words models) and find that it is largely driven by some simple heuristics at the word level that are ecologically valid in the SNLI dataset. This highlights the potential difficulties in training such general purpose function approximators such as RNNs for a task as complex and specific as understaning language, and the importance of having structured datasets in better understanding the performance of these systems (as well as in improving their performance).
Keywords: Sentence embeddings; compositionality; test data sets

Introduction and Motivation

In the world of deep learning, statistical structure is extracted by training on very large datasets. Several of these models are in theory Turing complete, leaving the discovery of the right parameters mostly to the training data and partially also to the training protocols. The training datasets are assumed to be samples from the natural distribution. While it remains an interesting question as to whether these training data sets truly are sampling the true distribution we want to approximate, another question is of how many random samples from the true distribution are enough to sufficiently constrain the space of models (in the huge space of possibilities the thousands of parameters in a neural network model allow) that explain the training data. Heuristics literature shows that simple strategies often satisfactorily explain the bulk of the variance amongst the most likely scenarios.

discussion on heuristics literature

Perhaps a better strategy for generating training data is to sample the cases that are most informative about the validity of competing models, especially to help rule out simple heuristics. This necessitates the disproportionate representation of 'edge cases' that do not have high probability of occurrence in the wild i.e. when sampling from the natural distribution. The creation of such "teaching" data sets better constrain the set of the possible models, instead of blindly sampling from the distribution posited by the true model.

discussion on "teaching data sets"

It is not a new proposal that one should focus on the boundaries when trying to learn structures, this has parallels to research on adversarial training. However, in the realms of very structured compositional domains like language, perhaps it is unrealistic to generate good adversarial examples that are valid. Enter the field of cognitive psychology and psycholinguistics where people have been coming up with good edge cases to test human understanding of these domains for decades.

discussion on what cogsci can offer / why adversarial training as it stands isn't there yet

The disproportionate sampling of tails of the true distribution, and the contrived cases that live there, is unlikely to occur in a natural environment. Perhaps kids are given some extra exposure to common pitfalls etc and what to avoid during schooling and other pedagogy, but it is disheartening to believe that structures cannot be learned from passive observation of natural statistics. However, generative learners are less sensitive to variations in the distributions of observed data and make better use of implicit negative evidence. So perhaps this kind of hand-holding is not necessary for more generative models.

Sentence Encoding

Understanding language requires the understanding not just of the words, but of their relation and ordering in a sentence. Models of word semantics where words are embedded in a vector space have been successful in capturing the meanings of words. However, due to the combinatorial productivity of language, the number of possible sentences far exceeds the size of the vocabulary, and generating similar vector embeddings for sentences has proved challenging. Recent literature reports several supervised as well as unsupervised approaches to learning sentence representations using Recurrent Neural Networks that account for word ordering (Kiros et al., 2015; Hill et al., 2016; Conneau et al., 2017). These are intended to capture their semantic content, and perform reasonably well on transfer tasks—i.e. other semantic tasks which the embeddings were not specifically trained on. Particularly, the performance of these sentence models on these tasks supersedes the performance of bag-of-words models that patently lack any relational information about the words, i.e. lack any

compositionality. However, it is unclear exactly what kind of compositional information is gained in these sentence models, above lexical meaning. We try to address this question by designing a dataset that relies on increasing degrees of compositional information, and is intractable for systems that capture only lexical information. Further, we try to characterize the compositional information in these sentence embeddings by understanding the cases where the sentence models succeed but the BOW models fail. Based on our observations, we hypothesize that the reason InferSent or modern neural network models fail to learn the compositionality of sentences is that they heavily rely on the data distribution, and overgeneralize patterns in training data. Given the exponentially large number of possible sentences, most subsets of sentences unless meticulously designed, can be justified by heuristics simpler than the true generative process of language. We analyze these heuristics and find their ecological validity in the SNLI dataset. Finally, we retrain a model on a dataset combining our ScrambleTest set and original SNLI train set to see if structured datasets like ours can be learned by the InferSent architecture.

NLI classifier

The sentence embeddings we use are from InferSent (Conneau et al., 2017). We choose to use these sentence embeddings as they hold 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 training set (Bowman et al., 2015). The training tasks 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. In the following, we use the pre-trained embeddings made available by the authors, featurized as suggested. We trained a classifier on the SNLI training set independently (using logistic regression as well as multilayer perceptron with a single hidden layer), with regularization optimized on the SNLI validation set. The bag of words (BOW) baseline model averages the GloVe embeddings for all the words in the sentence to form a sentence embedding. The BOW-LR and BOW-MLP achieve 50.24% and 53.99% accuracy respectively and the InferSent embeddings (LR and MLP) achieve 83.61% and 83.45% accuracy, on the SNLI test set (comparable to the performance in the end-to-end trained classifier as reported in Conneau et al. (2017)). Details of The logistic regressions are in Figures 2 and 3, MLP results look qualitatively similar.

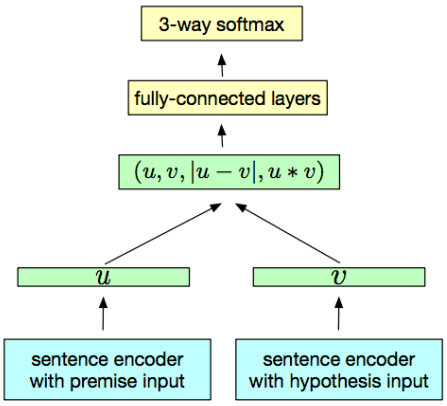

Figure 1: InferSent architecture. Source Conneau et al. (2017)

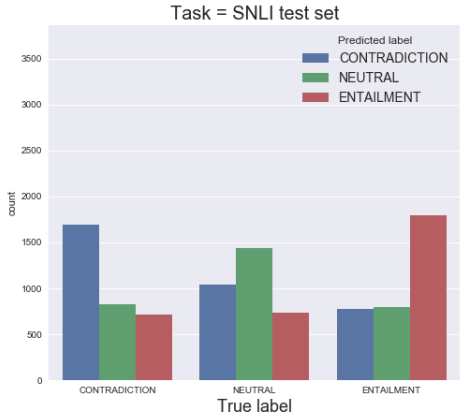

Figure 2: SNLI test Data: Logistic regression on BOW embeddings

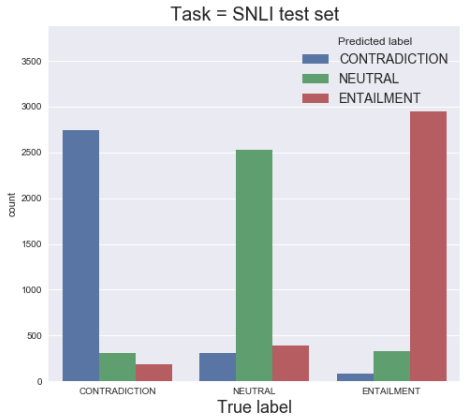

Figure 3: SNLI test Data: Logistic regression on InferSent embeddings

Type	Number of sentence pairs
Adjective-Referent binding	9600
Adjective-Referent binding (who is)	9600
Comparisons (more/less)	4800

Test data set

Our goal is to design sets of pairs of sentences such that the 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.

There are different levels of difficulty in the kinds of tasks we consider. The datasets are made up of two sets of sentence pairs, with two different relationships (of entailment, contradiction and neutral), such that the words in each sentence of the pair are the same across the set, but are reordered. Each data-set consists of such sets of sentence pairs, so that there are equal numbers of sentence pairs with two different relationships (among entailment, neutral and contrdiction).
Therefore, the maximum possible performance of a bag of words model is 50%, since it cannot distinguish the two pairs in the set, and cannot classify them differently. This provides a hard baseline for the performance that is possible without compositional understanding. Any performance above the BOW model is often seen as proof of compositionality. However, this is an unfair comparison the bag of words model baseline as implemented for comparsons receives only averaged word vectors for the sentence and therefore theoretically also loses some of the lexical information. We can gauge the compositionality that InferSent learns by seeing how differently it classifies these scrambled sets, and measuring its performance above 50%. Further, we make two copies of each dataset, one with exclusively short noun phrases (eg. 'the boy'), or long noun phrases (eg. 'the boy holding an umbrella') as a control to see how length affects performance.

Verb Symmetry

Some verbs are symmetric and some are not in terms of the subject and the object. We can form pairs of sentences by exchanging the subject and the object that have entailment, contradiction, or neutral relationships. We have 2 subcategories of these kinds of examples.

Contradiction type Where the word reordering changes the relationship from entailment to contradiction.

```
A: The woman overtakes the man
B: The woman overtakes the man
ENTAILMENT
A: The woman overtakes the man
B: The man overtakes the woman
CONTRADICTION
```

Neutral type Where the word reordering changes the relationship from entailment to neutral.

```
A: The woman watches the man
B: The woman watches the man
ENTAILMENT
A: The woman watches the man
B: The man watches the woman
NEUTRAL
```

Comparisons

When comparing two entities, the order of the entities matters in the sentence. We have 3 subcategories of comparisons.

More-Less type Where the A-B pairs in each pair of sentence 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 Where the A-B pairs in each pair of sentence 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
```

Same type Where the A-B pairs in each pair of sentence differ only in the order of the words.

```
A: The woman is more cheerful than the man
B: The woman is more cheerful than the man
ENTAILMENT
A: The woman is more cheerful than the man
B: The man is more cheerful than the woman
CONTRADICTION
```

Temporal ordering

Indication of the relative temporal ordering of events is also sensitive to the ordering of the same words in the sentence.

```
A: The woman stood up after the man stood up
B: The woman stood up after the man stood up
ENTAILMENT
A: The woman stood up after the man stood up
B: The man stood up after the woman stood up
CONTRADICTION
```

Adjective-Referent binding

To keep track of which noun in the sentence an adjective is referring to, the order of the words is critical.

```
A: The tall woman met the short man
B: The woman met the short man
ENTAILMENT
A: The tall woman met the short man
B: The short woman met the man
CONTRADICTION
```

With who-is We include a variant in which the adjective is bound to the referent with "who is".

```
A: The woman who is tall met the man who is
short
```

```
B: The woman met the man who is short
ENTAILMENT
A: The woman who is tall met the man who is
short
B: The woman who is short met the man
CONTRADICTION
```

Subject-Verb binding

In phrases where two subjects do different things, it is required that the relative relation between the verb and noun phrases be retained.

```
A: The woman stands up, however the man sits
down
B: The woman stands up.
ENTAILMENT
A: The man stands up, however the woman sits
down
B: The woman stands up.
CONTRADICTION
```

Negating a condition

There is unequal effect of negating the condition for a phenomena or the phenomena itself.

```
A: If there is a lot of snow, it is very cold
B: If there is a lot of snow, it is not very
cold
CONTRADICTION
A: If there is a lot of snow, it is very cold
B: If there is not a lot of snow, it is very
cold
NEUTRAL
```

Type		BOW		InferSent	
		LogReg	MLP	LogReg	MLP
Adj-ref	long	39.2	50	51.91	**51.39**
	short	49.71	50	53.25	**54.38**
Adj-Ref (who is)	long	37.04	49.83	65.81	**64.67**
	short	49.54	50	72.06	**72.34**
Comp (more/less)	long	0.08	7.38	43.10	**50.46**
	short	4.5	18.79	42.73	49.69
Comp (not)	long	3.15	45.38	42.54	39.38
	short	37.31	49.98	31.94	34.96
Comp (same)	long	12.44	50	50	**50.79**
	short	47.58	50	66.83	**79.81**
Neg-Cond		1.39	0	14.58	34.72
Subj–Verb	long	36.08	46.58	**53.17**	50.75
	short	47.25	50	53.79	**56.17**
TempOrd	long	17.5	50	50	50
	short	35.5	50	50	50
VerbSym (contr)	long	20	50	51	50.5
	short	45	50	63	**63.5**
VerbSym (neut)	long	50	50	50	50
	short	50	50	50	50

Experiments

We first do the analysis based on the percentage of correct classifications obtained by each of the 4 classifiers we have – MLP (with one hidden layer) trained on InferSent embeddings or trained on averaged GloVe embeddings, i.e. the Bag-of-Words baseline, as well as a logistic regression trained on each of these embeddings.

We see that, as expected, BOW never exceeds 50% precision. In fact, BOW thinks almost all of the pairs are entailments. The MLP performs better than the LogReg in general, although their performance on the SNLI test set was close to identical.

Classification Analysis

In order to better understand InferSent's performance, we look at the specific classifications it makes. We just consider the MLP results on short noun phrases. Resulst for long noun phrase, and both results for logistic regression results are in the appendix. We see that for BOW, the classifications made are exactly symmetric across the two true categories in each task (Figure 5) by design, since members of each category are just scrambled versions of each other and BOW cannot distinguish them. A sign of using more than word-level information is the asymmetry between the classifications of the two categories.

We divide the performance of InferSent into three categories 'high asymmetry, high scores' where more than word level information is used to make the right inferences, 'high asymmetry (Figure 6), low success' where more than word level information is used, but not to make the right inferences (Figure 7), 'low asymmetry, low success' where InferSent fails to use more than word level information (Figure 8). Here we designate 'high' perfromance very loosely to anywhere with more than 50% scores. We choose this set of distinctions because not using more than word level information will ensure bad performance on these tasks, but use of more than word level information doesn't guarantee good performance. This is because each task consists of members of two classes, but there are 3 classes into which the classifier can place each pair. Classification into the third class will always result in low performance, even if the classification differentially uses more than word level information (i.e. if there is high asymmetry).

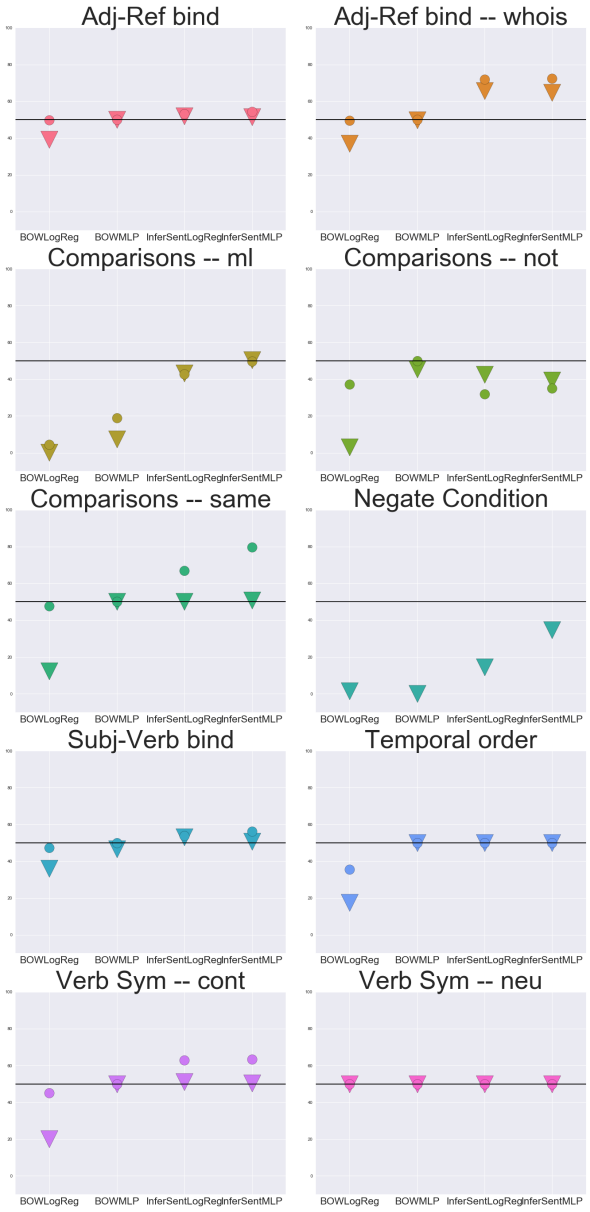

Figure 4: Performance on all tasks in ScrambleTest. Triangles represent 'long' noun phrases, circles represent 'short' noun phrases

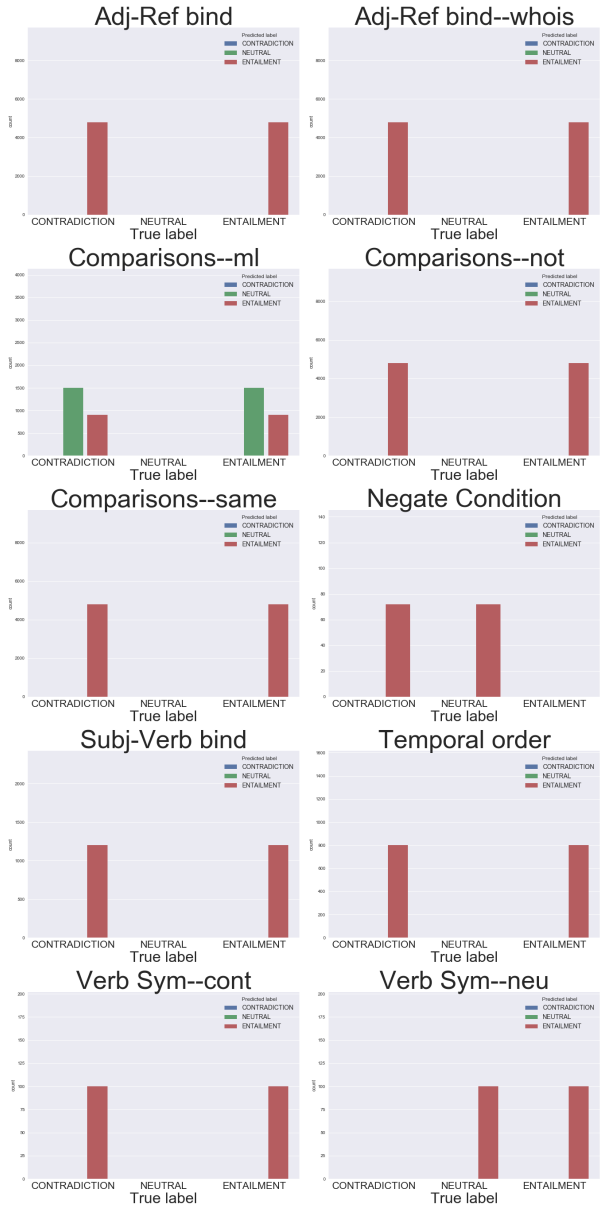

Figure 5: The classifications made by the MLP on the BOW vectors. The symmetries across the classifications are clear.

Observations

We analyse some of the patterns we observe and study the possible origins of this behavior based on the statstics of the SNLI dataset.

All same words When the words in both sentences are the same, they are classified as entailing one another, except in the presence of certain words. We observe that in SNLI dataset, most contradictory sentence pairs have no overlap in words. It's much more likely for a sentence pair to be entailment or neutral if they have a lot of words overlap. 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.`
`CONTRADICTION`

In order to qualitatively verify this observation, we rank all the sentence pairs by overlap rate: $\frac{\#\text{ of overlap words}}{\text{total }\#\text{ of words}}$ (in non-increasing order). We then look at top X sentences with highest overlap (or *relevance*), and we observe:

As shown in Table 1, 91.5% of the pairs with maximum overlap between the sentences have the true label of either entailment or neutral, and are very rarely labelled contradiction.

Difference of one word When the words in two sentences differ by just one word, the decision is largely based on if those words have opposing meanings irrespective of the or-

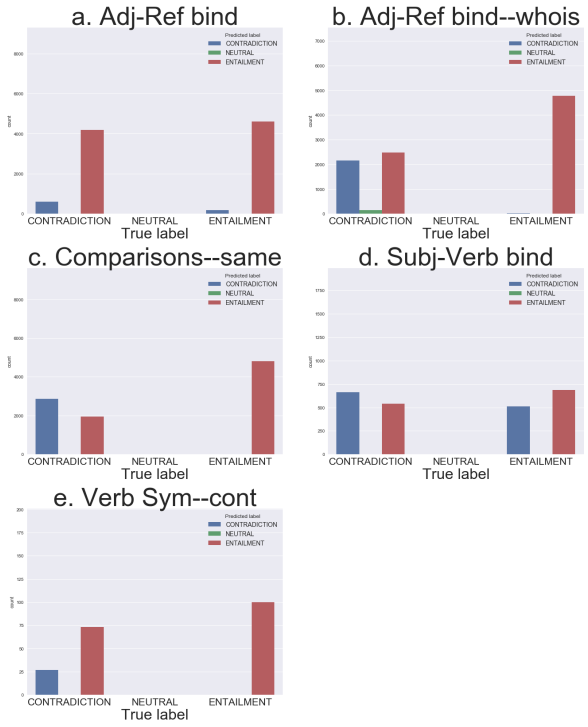

Figure 6: Classifications by the MLP on the InferSent vectors with short noun phrases, tasks with 'high asymmetry, high scores'.

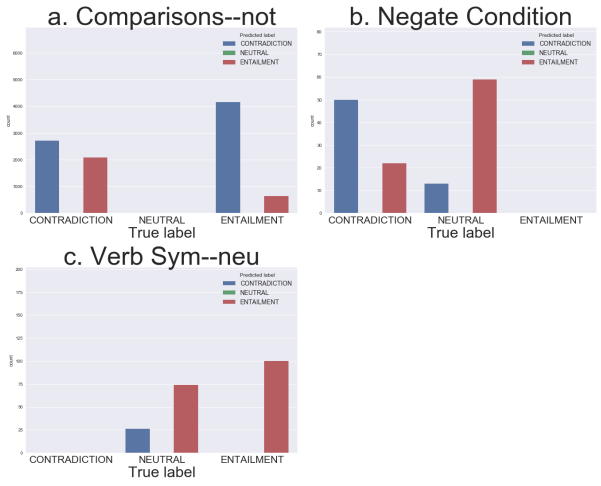

Figure 7: Classifications by the MLP on the InferSent vectors with short noun phrases, tasks with 'high asymmetry, low scores'.

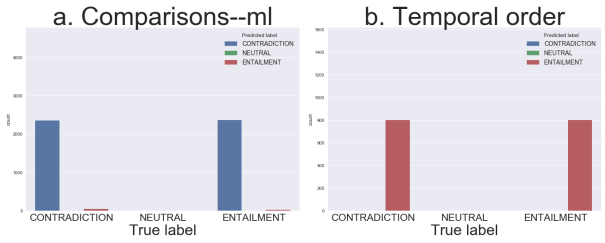

Figure 8: Classifications by the MLP on the InferSent vectors with short noun phrases, tasks with 'low asymmetry, low scores'.

Top	Entailment	Neutral	Contradiction
All	183416 (33.4 %)	182764 (33.3 %)	183187 (33.3 %)
10000	3954 (39.5 %)	3567 (35.7 %)	2479 (24.8 %)
1000	508 (50.8 %)	407 (40.7 %)	85 (8.5 %)

Table 1: High Overlap of words in SNLI

der of the words. We see this from performance on more-less type comparisons (Figure 8a). Here the words across the pairs differ only in the presence or absence of the word 'more' or 'less'. For example:

```
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
```

Since the relation between the words 'more' and 'less' is largely contradictory, their use in pair of sentences leads the classifier to presume the sentences are contradictory, irrespective of the order of the words.

We now observe that this heuristic is fairly consistent with SNLI data.

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

Table 2: Antonym word pair in the SNLI data set

We see in Table 2 that the presence of antonyms strongly predicts a true label of contradiction in the SNLI data set. If we consider only the top 1000 in the high overlap set, the trend is more evident with 43.5% contradiction sentence pairs containing antonyms, and only 8.7% entailment sentence pairs containing antonyms.

Negations We see that comparatives that differ in the presence or absence of the negation 'not', are largely classified as contradictions (Figure 7a). To verify that our observation strongly correlates with SNLI train data, we look at sentence pairs that contain "negating ngrams": no, not, n't, don't, doesn't.

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

Table 3: Overall Data Negation statistics

We see in Table 3 that the presence of negations strongly predicts a true label of contradiction in the SNLI data set. If we consider only the top x in the high overlap set, the trend is more evident. We observe that 60.0% sentence pairs with

negation are contradictions, and only 7.5% are entailment.

The effect of permuting So far we have focused on word level information in the statistics of SNLI in explaining much of InferSent's behavior. A systematic analysis of the effect of word order if challenging due to the combinatorial explosion in the number of possibilities. But we make a few observations.

In some cases, the order of words has no effect whatsoever on the classification made by InferSent, for example with temporal ordering type sentences (Figure 8b).

```
A: The woman stood up after the man stood up
B: The woman stood up after the man stood up
ENTAILMENT
A: The woman stood up after the man stood up
B: The man stood up after the woman stood up
CONTRADICTION
```

This is perhaps because the SNLI dataset does not contain many time-ordered events, and has not learned the meanings of the words 'before' and 'after' and that they promote ordering, therefore defaulting to entailment due to the high overlap.

The presence of words/short N-grams that are known to contradict each other ('more X'/'less X', 'is'/'is not', other antonyms), or just words that indicate order sensitivity ('more', 'overtake'), seem to elicit some word order sensitivity for InferSent. From qualitative analysis, we observe that if common antonyms are swapped in order, then Infersent will likely classify this sentence pair as a contradiction. For example:

```
A: Boy sitting with a girl on the bench in the park
B: Girl sitting with a boy on the bench in the park
```

We conjecture that simply a larger perturbation to the permutation of words (in the presence of these "order promoting" words) takes the classification closer to contradiction. This is trivially true for cases where zero perturbation results in an entailment inference, and a perturbation to order the order sometimes leads to a contradiction (same-type comparatives, verb-symmetries in Figures 6c,e 7c). But the key observation that supports this conjecture is in the case of comparatives of the 'not' type 7a.

```
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
```

Here, all pairs of sentences differ in the presence of 'not', and as observed in the section on 'negations' we see that all of these are more likely classified as contradictions than entailments. But in this dataset, pairs that more perturbed in the word order are in fact entailments.

We see the reverse trend in the classifications made by InferSent sentences that are truly entailments, and are more perturbed in word order, are more likely to be classified as contradictions than true contradictions. Further, we also find that while all comparatives of the 'more-less' type are classified as contradictions (Figure 8a), the system is more confident about the ones that are truly entailment (1867/2400 pairs) being contradictions, i.e. the ones that have a larger perturbation to word order.

Other asymmetries InferSent perhaps encodes meanings of some ngrams. This is supported by asymmetry in subject-verb binding(Figure 6d)

```
A: The woman stands up, however the man sits
down
B: The woman stands up.
ENTAILMENT
A: The man stands up, however the woman sits
down
B: The woman stands up.
CONTRADICTION
```

Where the distinction requires the system to encode the binding of the words 'woman sits' and 'man stands'. It also shows asymmetry on adjective-reference binding(Figure 6a)

```
A: The tall woman met the short man
B: The woman met the short man
ENTAILMENT
A: The tall woman met the short man
B: The short woman met the man
CONTRADICTION
```

Where the binding for the word pairs 'tall woman' and 'short woman' are important. However, the performances on both of these is very low.

Also, perhaps the combining of a phrase with 'not' is encoded. This could be argued from the asymmetry in negating a conditional (Figure 7b).

```
A: If there is a lot of snow, it is very cold
B: If there is a lot of snow, it is not very
cold
CONTRADICTION
A: If there is a lot of snow, it is very cold
B: If there is not a lot of snow, it is very
cold
NEUTRAL
```

Where the binding of not with the different verb-phrases would allow for the observed asymmetry. However, the number of sentence pairs in this dataset is fairly low and is more difficult to draw conclusions from

SNLI test set

Our ScrambleTest set is not balanced, and was not part of the training. For a more controlled way to understand the differences between the performance of InferSent and Bow, we try to also look at the cases in the SNLI test set that are correctly classified by InferSent but incorrectly classified by BOW, with a high margin/confidence. See Table 4 for examples. It seems from these that BOW learn contradiction only as 'no overlap in words' whereas although InferSent is also

Data	Type	Sentences	True	IS-LR	BOW-LR
SNLI data	Test set	A: A runner in a black and blue uniform competes in a race. B: he is winning.	neut	neut (99.99%)	contr (99.89%)
		A: A boy runs as others play on a homemade slip and slide. B: A boy is running.	entail	entail (99.95%)	contr (99.89%)
		A: A man sits at a table in a room. B: A woman sits.	contr	contr (99.99%)	entail (99.75%)

Table 4: High Margin misclassifications by BOW

strongly affected by the extent of overlap in words, it is able to have a more nuanced encoding ('man sits' vs 'woman sits'). This also explains why BOW classifies almost all of the pairs in our data set as entailments, since most of the pairs in our dataset have a high overlap in the words used. Further, it also is able to encode meanings that straddle two or more words, for example that 'runs' entails 'is running'. This is consistent on what we observed about bi-grams above. However, it is difficult to make general claims.

Augmented training

Now, we experiment with retraining the InferSent model on a combined dataset which includes both new ScrambleTest and original SNLI training data.

The new ScrambleTest data uses a similar generation process as our old ScrambleTest data. However, we only generate data of the 'comparisons' type since the performance of InferSent on these is the most interpretable. The training dataste is generated as follows: we collect positive comparison phrases (p) such as "more adj." where "adj." is some adjective, and negative comparison phrases (n) such as "less adj.". The two sentences in a pair are of the form "(np1) is (p) than (np2)." and "(np2) is (n) than (np1).", where (np1) and (np2) are applicable noun phrases. The adjectives and noun phrases are randomly sampled from a large set of given possibilities. In total, we generated 40k such sentence pairs, which is roughly 7% of overall SNLI training data (550k sentence pairs).

Moreover, in order to minimize the effect of other factors such as vocabulary distribution. We intentionally modified our generation process, so that the vocabulary distribution of this newly generated ScrambleTest is similar to original SNLI training dataset. Only 7 words differ by more than 1% in their occurrence in SNLI: not, a, than, the, is, less, more.

Summary of findings

The classifier based on InferSent definitely does not capture all of the compositionality in sentence structure. While it is difficult to pin down exactly what it is that the classifier does pick up on, certain behaviors draw attention to some potentially heuristic encodings of sentence compositionality.

First, we see that the difference of one word between the two sentences, when the difference is an antonym or a negation, leads to InferSent tending to classify them as contradictions irrespective of word order. This is best highlighted in the more/less type and not type comparison dataset tests. This heuristic has ecological validity in the training set as shown in Tables 2 and 3, we see that the presence of negations and antonyms strongly correlate with a contradiction label. This illustrates a disproportionate dependence on lexical, rather than compositional meaning in InferSent.

Second, in cases where even the words across the pairs are the same, we see that sometimes the order of the words is detected, and taken into account for the classification (as indicated in performance on the verb-symmetries and the same-type comparisons). This is sometimes detected and (correctly in this case) tagged a contradiction. However, for the not-type comparisons, the more jumbled pairs are in fact entailments. The classifier does the opposite, with true entailments (that is more jumbled ones) more often being classified as contradiction than the true contradictions. This hints at one way in which order information might be used, but is a heuristic that might often work. Finding ecological valididty for permutations in SNLI is challenging and we leave this to future work.

Finally, we see that there are some other cases in which InferSent performs above 50%, indicating that it is picking up some information about the relative positions of the words, although clearly not enough to perform competitively on these compositional tasks. Further work is needed to better isolate exactly what is learned, but is beyond the scope of this work.

This work demonstrates the inadequacy of most datasets available today in truly testing if compositional structure, i.e. structure beyond lexical structure, is being picked up by NLP models. InferSent achieves very high performance on the test set of the SNLI dataset, *as well as several other tasks*, but fails on our dataset. Our ScrambleTest dataset is available online to test future models on. This work highlights the value of crafting diagnostic datasets, in the spirit of tasks designed in in cognitive psychology, in better understanding the behavior of these complex models. Our results on augmented training also show promise for systematic training sets... xyz.

Future Directions

This is one step towards building better diagnostic test sets to score NLI models like InferSent. Other interesting things to try would be to see how other models, particularly generative models like SkipThought, differ on these metrics. Further, we could use more sophisticated methods like LIME (Local Interpretable Model-Agnostic Explanations, Ribeiro et al. (2016)) to generate interpretable explanations for its classifications and understand the performance of these models on diagnostic datasets.

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.

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

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R., & Fidler, S. (2015). Skip-Thought Vectors.

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.

Other Figures

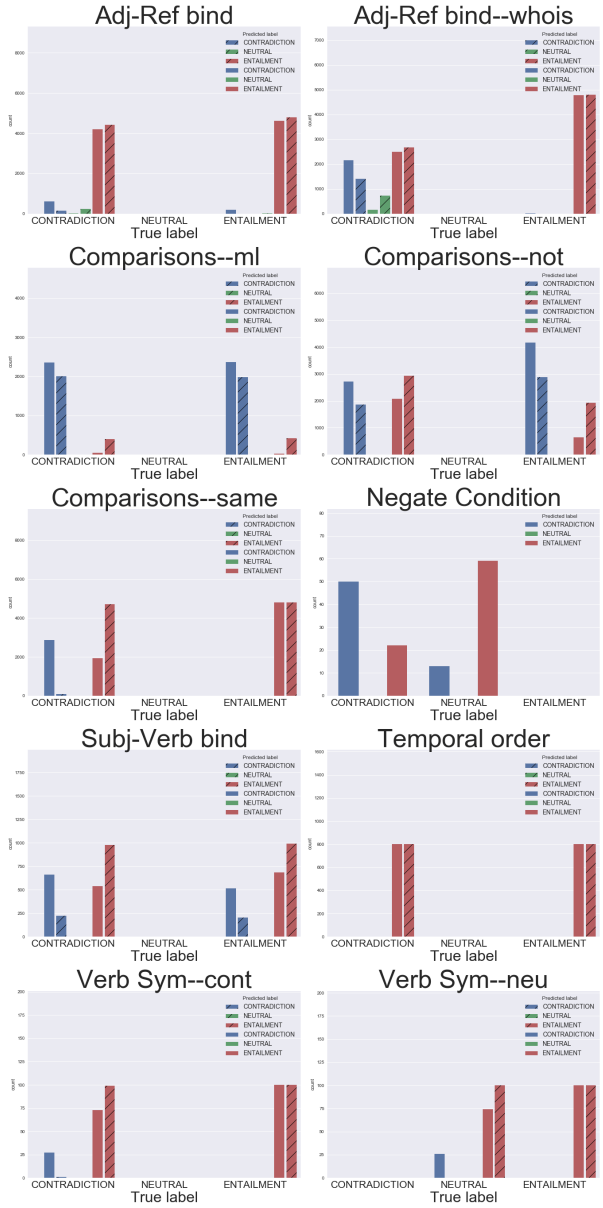

Figure 9: ScrambleTest Data: The classifications made by the MLP on the InferSent vectors. Long noun phrases are hatched.

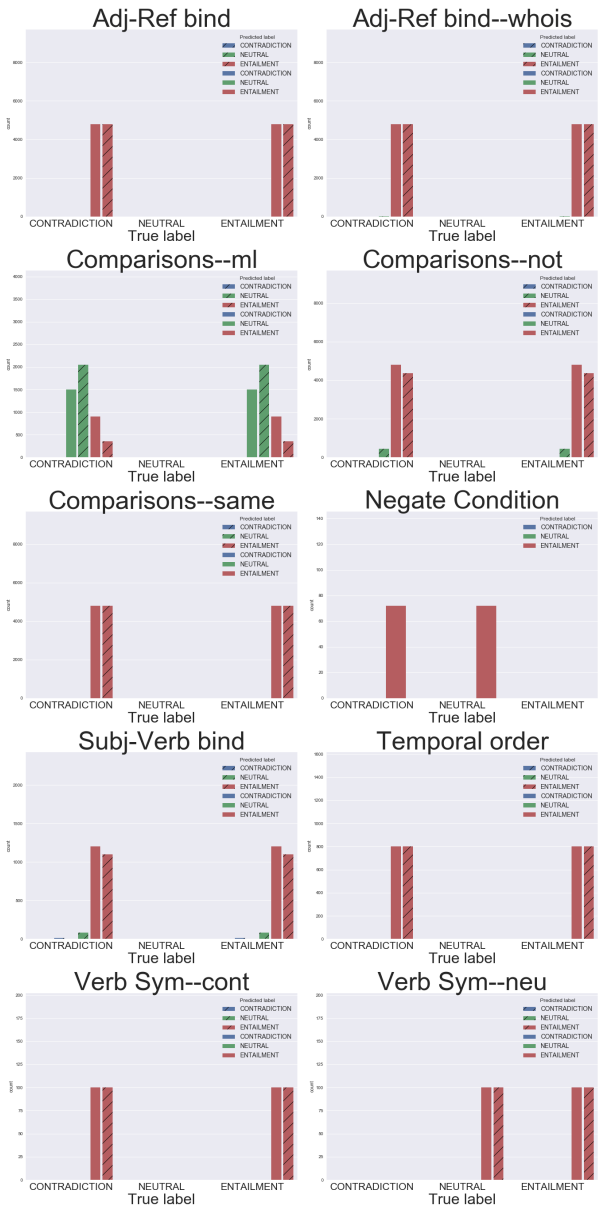

Figure 10: ScrambleTest Data: The classifications made by the MLP on the BOW vectors. Long noun phrases are hatched. Note that judgments are symmetric for both kinds of true labels, by design.

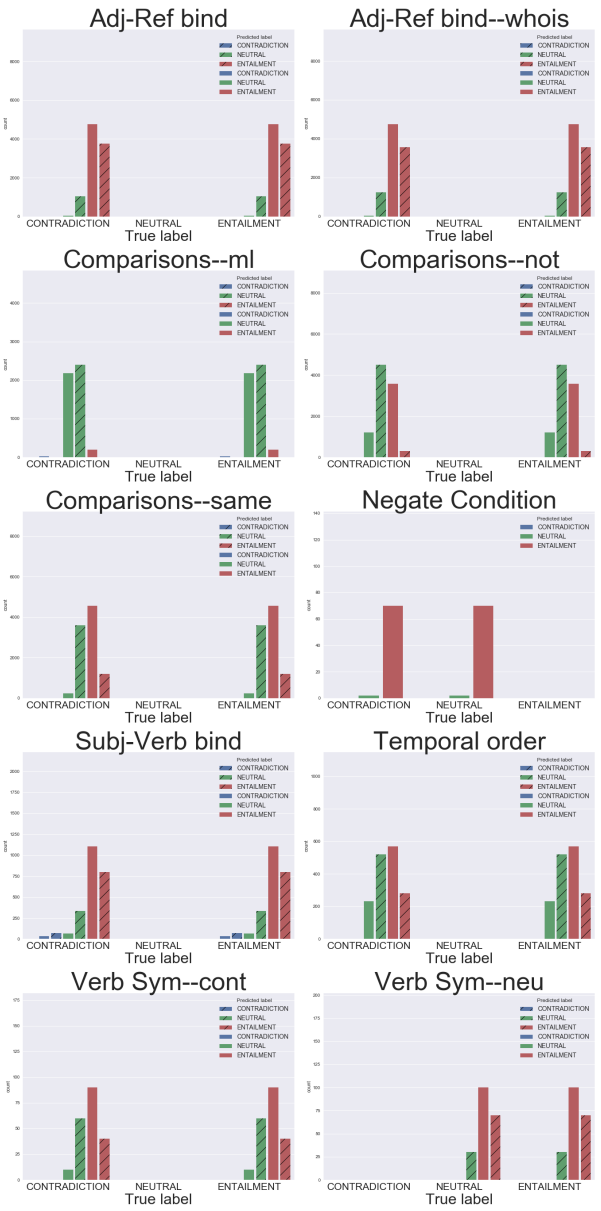

Figure 11: ScrambleTest Data: The classifications made by the logistic regression on the BOW vectors. Long noun phrases are hatched.

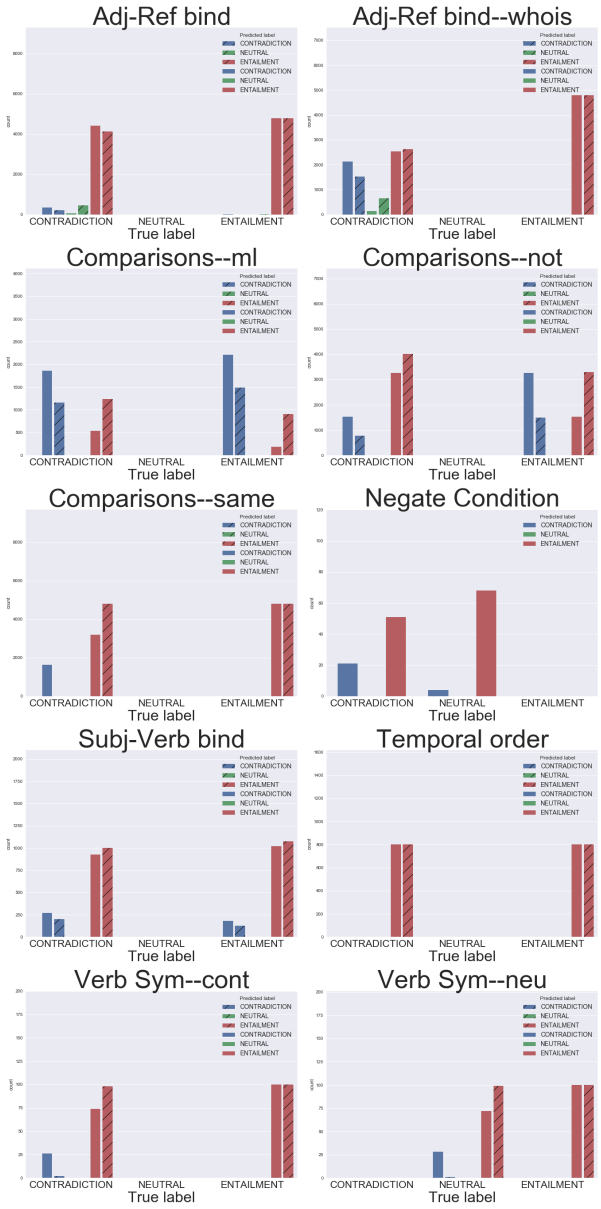

Figure 12: ScrambleTest Data: The classifications made by the logistic regression on the InferSent vectors. Long noun phrases are hatched.