Gauging Compositionality in Sentence Embeddings

Ishita Dasgupta[1], **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

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. While these sentence representations 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 (**?**), 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 InferSent encoding some bi and tri grams that occur with high frequency, and the ability to detect permutations of word order, rather than encoding truly sentence-level representations.o

Keywords: Sentence embeddings; compositionality; test data sets

Introduction

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 (**???**). 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 in increasing degree 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. [demi: Based on our observations, we hypothesize that the reason InferSent or modern neural network models, especially discriminative ones, failed to learn the compositionality of sentences is that they heavily relies on the data distribution, and often tends to overgeneralize patterns in training data. Thus, data is essential to learn a model with compositionality. We attempt to verify our hypothesis by further analyzing some of our previous observations to understand how the model learned to behave in this way. Finally, we retrained a model on a dataset combining our carefully designed ScrambleTest set and original SNLI train set.]

NLI classifier

The sentence embeddings we use are from InferSent (**?**), the current state of the art in performance on transfer in semantic tasks. We expect that this indicates that the embeddings provide a good representation for the semantics of a sentence. These embeddings were trained end-to-end using the architecture in Figure **??** on the SNLI training set (**?**). The task was 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. Details of The logistic regressions are in Figures **??** and **??**, MLP results look qualitatively similar.

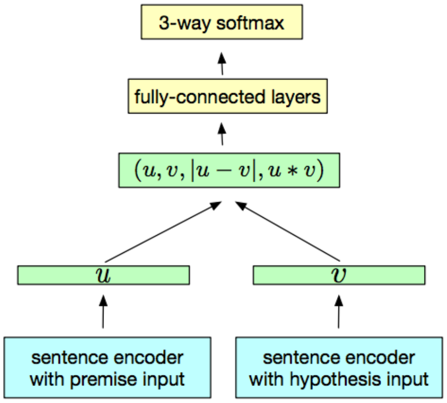
3-way softmax
fully-connected layers
$(u, v, |u - v|, u * v)$
u
v
sentence encoder with premise input
sentence encoder with hypothesis input

Figure 1: InferSent architecture. Source **?**

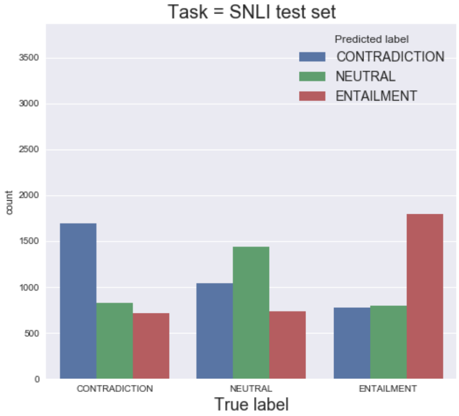

Figure 2: SNLI test Data: Logistic regression on BOW embeddings

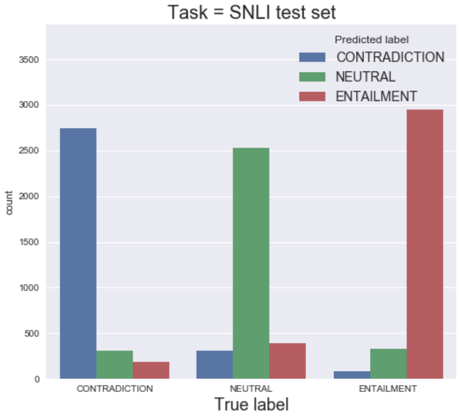

Figure 3: SNLI test Data: Logistic regression on InferSent embeddings

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 pairs with one of two relationships (among the 3 possible of entailment, neutral and contrdiction) in equal number.

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 receives only averaged word vectors for the whole sentence. That is, it doesn't even receive all 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

Experiments

We first do the analysis based on the percentage of correct classifications obtained by each of the 4 classifiers we have – logistic regression trained on InferSent embeddings or trained on averaged GloVe embeddings, i.e. the Bag-of-Words baseline, as well as an MLP (with one hidden layer) 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 for now, long noun phrase results, and logistic regression results are in the appendix. We see that for BOW, the labels are exactly symmetric across the two true categories in each task (Figure **??**) by design, since members of each category are just scrambled versions of each other and BOW cannot distinguish them. A

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

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 **??**), low success’ where more than word level information is used, but not to make the right inferences (Figure **??**), ‘low asymmetry, low success’ where InferSent fails to use more than word level information (Figure **??**). Here we designate ‘high’ perfromance to anywhere with more than 50% scores. We choose this distinction 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).

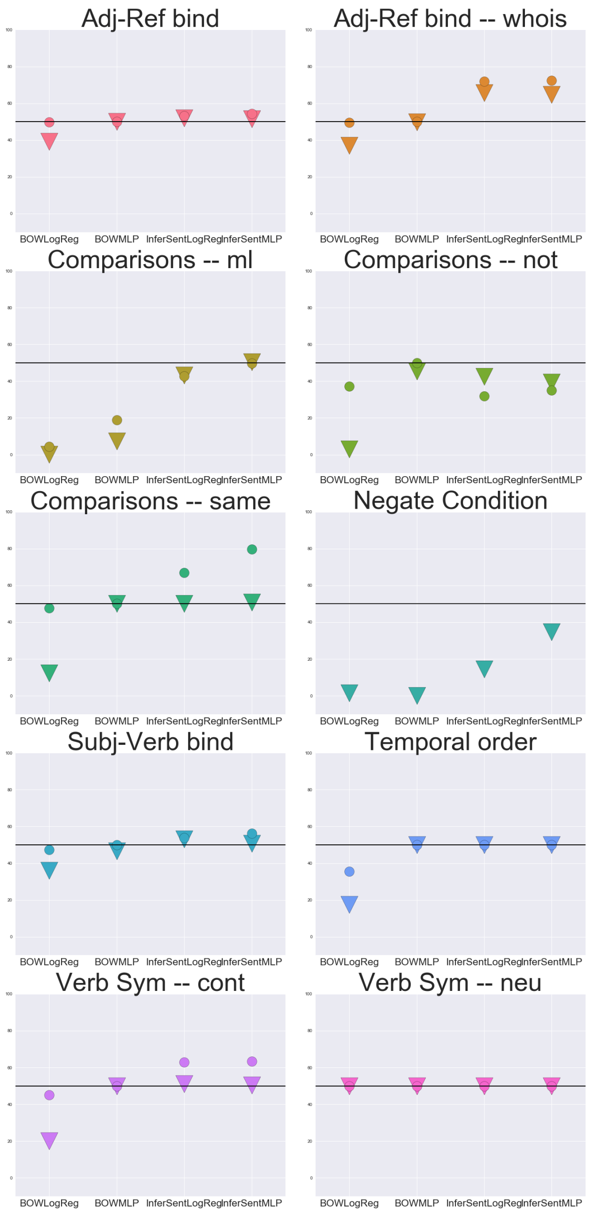

Figure 4: Performance on all tasks in ScrambleTest. Triangles represent ‘long’ noun phrases, circles represent ‘short’ noun phrases

Observations

All same words When the words in both sentences are the same, they are classified as entailing one another, unless there is presence of certain words that have been learned to promote ordering sensitivity. We see that the order of the sentence leads to no difference in classification of temporal ordering type sentences (Figure **??**b).

```
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

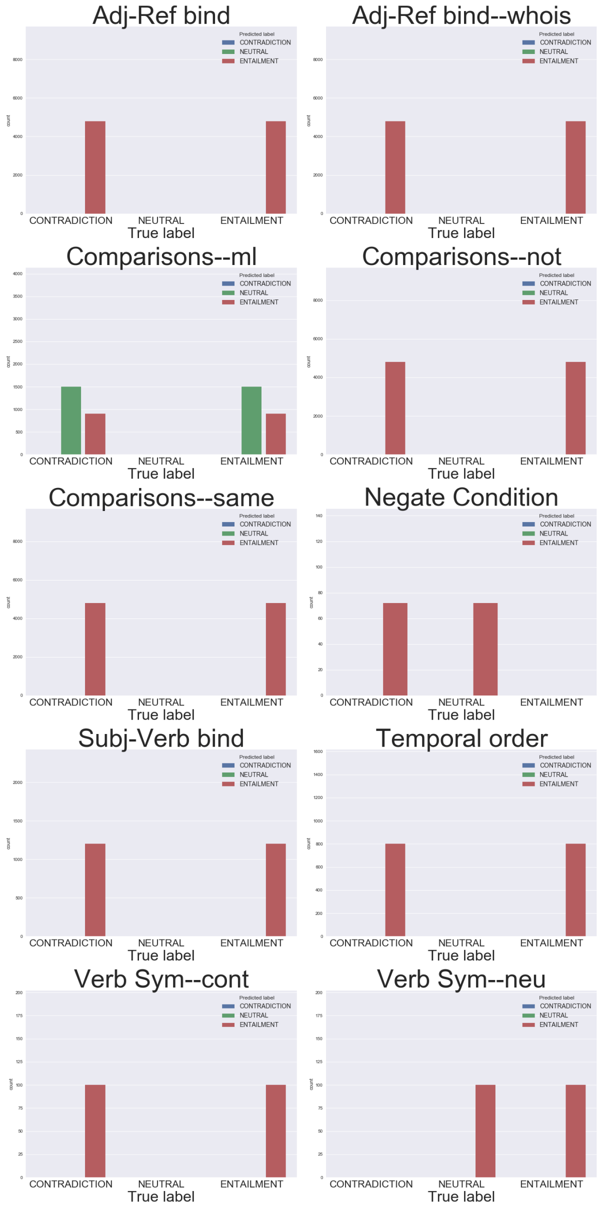

Figure 5: The classifications made by the MLP on the BOW vectors. The symmetries across the classifications are clear.

many time-ordered events, and has not learned the meanings of the words 'before' and 'after' and that they promote ordering. However, other sentence pairs in which the words remain the same across the pair, like the sentence pairs for same-type comparisons that contain words like 'more' and 'less'(Figure **??**c), as well as the verb symmetry tasks that contain verbs like 'overtake' (Figures **??**e & **??**c) do demonstrate some order sensitivity.

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 order of the words. We see this from performance on moreless type comparisons (Figure **??**a). Here the words across

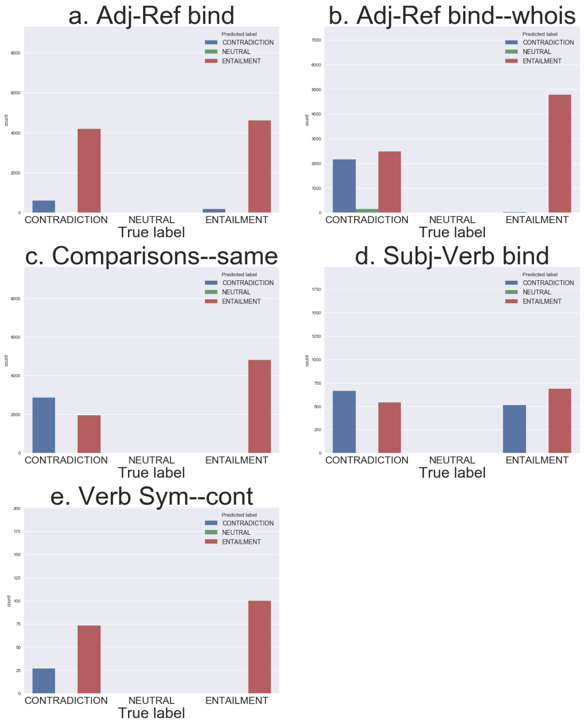

Figure 6: Classifications by the MLP on the InferSent vectors with short noun phrases, tasks with 'high asymmetry, high scores'.

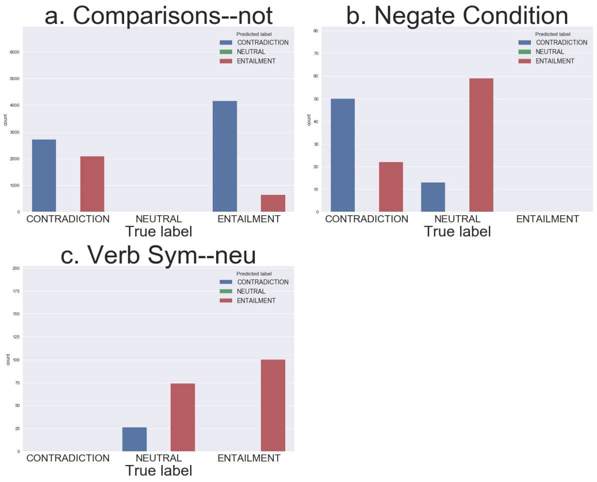

Figure 7: Classifications by the MLP on the InferSent vectors with short noun phrases, tasks with 'high asymmetry, low scores'.

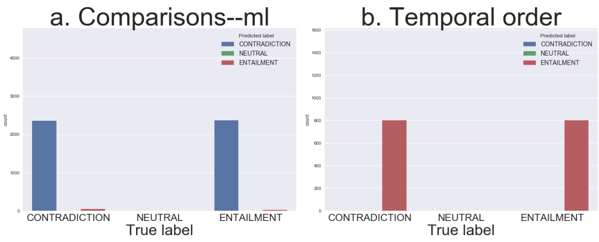

Figure 8: Classifications by the MLP on the InferSent vectors with short noun phrases, tasks with 'low asymmetry, low scores'.

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. Similarly, though less clearly, we see that comparatives that differ in the presence or absence of the negation 'not', are largely classified as contradictions (Figure **??**a).

Negating bi-grams Negating bi-grams, i.e. combining a phrase with 'not' is perhaps encoded. This could be argued from the asymmetry in negating a conditional (Figure **??**b).

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, it could also be explained by sensitivity to where in the sentence the 'not' occurs, with the first 'is' or the second. Performance on this tasks stays low. despite high asymmetry, due to misclassification of all the neutral pairs in the dataset as entailment.

Other bi-grams InferSent perhaps encodes meanings of some other bigrams. This is supported by asymmetry in subject-verb binding(Figure **??**d)

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 **??**a)

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.

The effect of permuting In the presence of words/short N-grams that are known to contradict each other ('more X'/'less X', 'is'/'is not', 'tall woman'/'short woman'), or just words that indicate order sensitivity ('more', 'overtake'), we can conjecture that it is simply larger perturbation to the permutation of 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 **??**c,e **??**c). But the key observation that supports this point is in the case of comparatives of the 'not' type **??**a. Here, all pairs of sentences differ in the presence of 'not', but those that more perturbed are in fact entailments.

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

We see the reverse trend in the classifications made by InferSent sentences that are entailments 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 **??**a), 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 order. This could also perhaps explain why the 'who-is' version of the adjective reference binding (Figure **??**b) shows a stronger effect than without (Figure **??**a), since one entails a larger perturbation to the order of the words because of the extra words 'who is'.

[demi: ?]

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 **??** for examples. It seems from these that BOW learn contradiction only as 'no overlap in words' whereas InferSent 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. A better understanding of what InferSent could achieve if also trained on some sen-

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 1: High Margin misclassifications by BOW

tences from our ScrambleTests dataset that highlight the value of word-order and compositionality, is left to future work.

[demi:

Hypothesis and Solution: Right Data Distribution

We hypothesize that the lack of compositionality of Infersent model comes from the misleading SNLI training data. Specifically, for each category misclassification of InferSent model (shown in **Observation** section), we attempt to explain InferSent model behavior by analyzing training data distribution.]

[demi:

All same words

We observe that in SNLI dataset, most contradictory sentence pairs are irrelevant sentences. It's much more likely for a sentence pair to be entailment or neutral if they have a lot of words overlap. If this is true, it could possibly explain why our model tends to classify similar sentences as entailment. For example, a contradictory sentence pair could be:
A:Several people are trying to climb a ladder in a tree .
B:People are watching a ball game .
In order to qualitatively verify this observation, we rank all the sentence pairs by overlap rate: $\frac{\text{\# of overlap words}}{\text{total \# of words in both sentences combined}}$(in non-increasing order). We then look at top X sentences with highest overlap (or *relevance*), and we observe:]

[demi: As shown in the table, when there are more overlap between two sentences, it's more likely to be entailment or neutral and less likely to be contradiction. Notably, there are only 8.5% contradictory sentence pairs in top 1000 overlap data.]

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 2: High Overlap Data Statistics

[demi:

Difference of one word

In previous section, we argue that our model tends to classify a sentence pair as "contradiction" if differ only 1 or a few words containing an antonym pair. We further confirm this observation, we see that 60.0% sentence pairs that Infersent classified as contradiction contain antonyms.
We now observe that it's consistent with SNLI data.In top 1000 overlap data, 43.5% contradiction sentence pairs contain antonyms, whereas only 8.7% entailment sentence pairs contain antonyms. Even when we include contradiction sentence pairs which are simply irrelevant, there are 12.2% contradiction sentence pairs contain antonyms, and 3.5% entailment sentence pairs contain antonyms. Overall, 61.2% sentence pairs containing antonyms are contradictions, and only 18.0% are entailment.
On the other hand, we found most popular antonyms in contradiction pairs in SNLI are: "women/men", "boy/girl", "black/white" ...... It's extremely consistent with the most popular antonyms in pairs that Infersent classified as contradiction.
Moreover, we qualitatively look at Infersent classification after changing only one word of a sentence. Consider two sentences:
A:"X driving a motorcycle with a girl sitting on back"
B:"Y driving a motorcycle with a girl sitting on back"
We see that when X/Y are popular antonyms such as man/woman, boy/girl, black/white, our model predicts contradiction. However, when X/Y are other antonyms such as old/young, sitting/standing, it still gives entailment.] [demi:

Negating

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.
On overall data, we have:
Here, we denote N as negation, E as entailment and C as

P(N|C)	P(C|N)	P(N|E)	P(E|N)
3.3%	58.4 %	1.1 %	20.0%

Table 3: Overall Data Negation statistics

contradiction. As we can see, negation correlates more with contradictions than entailment almost 3 times.

Moreover, in top 10,000 overlap data (relevant), in SNLI, we observe that 60.0% sentence pairs with negation are contradictions, and only 7.5% are entailment.

Note that we also did same experiment on more broadly defined negation ngrams which includes Neither, None and etc, and we get similar results. We have 43.6% negation pairs are contradiction, while 27.3% negation pairs are entailment. 7% contradiction pairs have negation while only 4% entailment pairs have negation.] [demi:

Permuting

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"
Moreover, we observe that when the order of common antonyms are swapped, then it's highly likely this is a contradiction in SNLI training dataset. (TODO:verify using quantitative data that my above claims are correct)]

[demi:

Retraining

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 "comparative" data. We have positive comparative phrase (p) such as "more adj." and negative comparative phrase (n) such as "less adj.". The two sentences will be in the format of such as "np1 is p than np2." and "np2 is n than np1".We give our data generator a large set of possible adjectives and noun phrases.

Then, when generating a data point, we randomly sample an adjective and 2 noun phrases. 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 (through picking noun phrases and adjectives and designing sample distribution), so that the vocabulary distribution of this newly generated ScrambleTest is similar to original SNLI training dataset. Only 7 words differ by at least 1% occurrence: not, a, than, the, is, less, more. Those are words which are necessary to have a difference.
]

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 patently misguided encoding of compositionality. We draw attention to two specific sub-tasks to highlight this.

First, we see in the more/less-type comparison class, where the words in the pairs of sentences differ only in whether they contain the word ‘more’ or the word ‘less’, that the presence of words that contradict each other in the two sentences of a pair consistently leads the classifier to think the sentences overall contradict each other. This illustrates a disproportionate dependence on lexical, rather than compositional meaning. The large percentage of the not-type comparisons, where the words in the pairs of sentences differ only in whether they contain the word ‘not’ or not, being classified as contradictions also support this finding.

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). Here, the same sentence repeated is the entailment, and a change of the order of the words indicates a contradiction. 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 patterns also holds for more/less-type comparisons, where the classifier is more confident about true entailments (i.e. the more jumbled pair) being contradictions, than the true contradictions (78% of the time). This hints at one way in which order information might be used, but just as a heuristic that might often work, but isn't fully utilizing compositionality.

Finally, we see that there are some other cases in which InferSent performs above 50%, indicating that it is picking up some more information, 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 also 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. Some of the sub-types, particularly comparisons, adjective-referent binding, and subject-verb binding have enough sentence pairs to be able to consider including some of them during training to augment the classifier. We leave this to future work.

Future Directions

- How much info do the sentence representations actually have?
  - BOW: How much is lost due to averaging, and to what extent is it possible to back out the words? Include ‘product of all word reps’ and other (?) operations that

    are symmetric to word position to improve info transfer about words in the sentence as a **fairer BOW baseline comparison**.
  - InferSent: Could a decoder recreate the sentence from the last hidden state of an RNN?

- Are the ScrambleTests examples just too different from SNLI to make a fair comparison?
  - Train on part of the ScrambleTest data, ensuring no overfitting and retest.
  - Ensure that distributional characteristics (base rates of words for eg, and other statistics) are comparable to SNLI.

- What are the differences between the results from LogReg and MLP?

- Test other sentence representations

- Find more precisely:
  - Description of incremental compositional information needed and how different types of scrambled data sets provide a paving.
  - What exactly InferSent classifies based on (LIME?). But perhaps that's getting too invested in InferSent in particular.

Other Figures

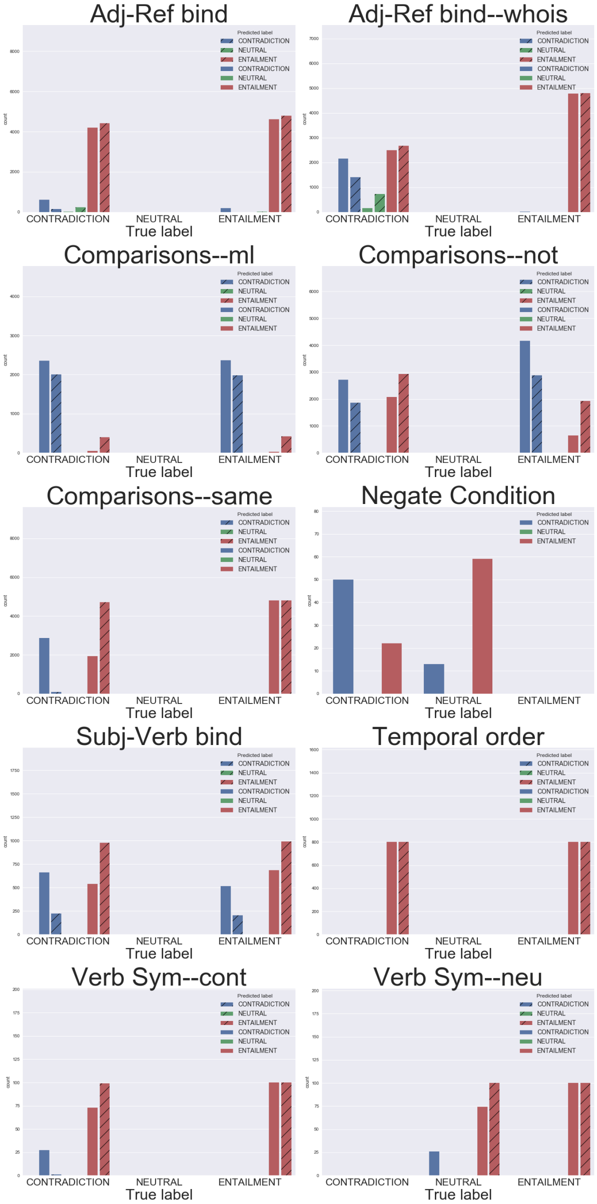

Figure 9: ScrambleTest Data: The classifications made by the MLP on the InferSent vectors. Long noun phrases are hatched.

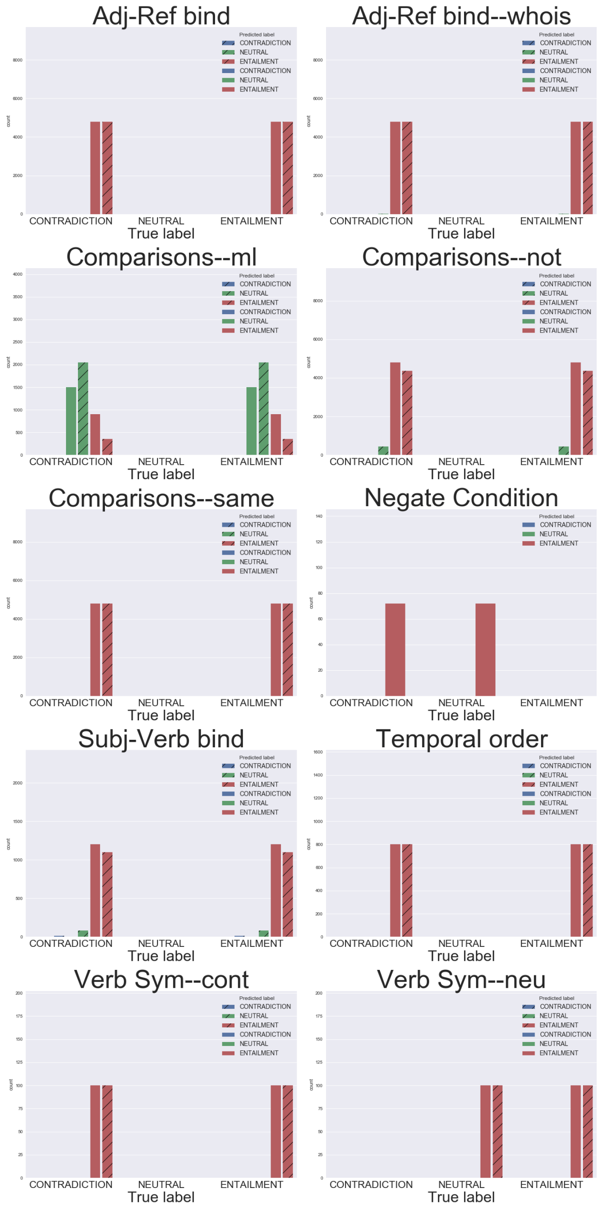

Figure 10: ScrambleTest Data: The classifications made by the MLP on the BOW vectors. Long noun phrases are hatched. Note that judgments are symmetric for both kinds of true labels, by design.

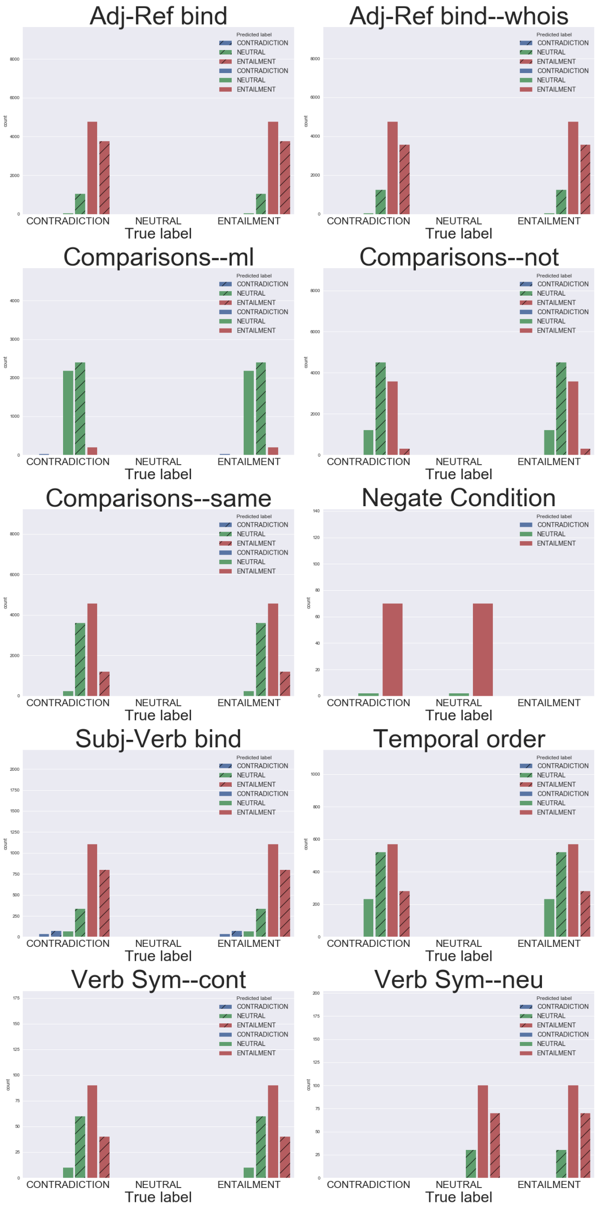

Figure 11: ScrambleTest Data: The classifications made by the logistic regression on the BOW vectors. Long noun phrases are hatched.

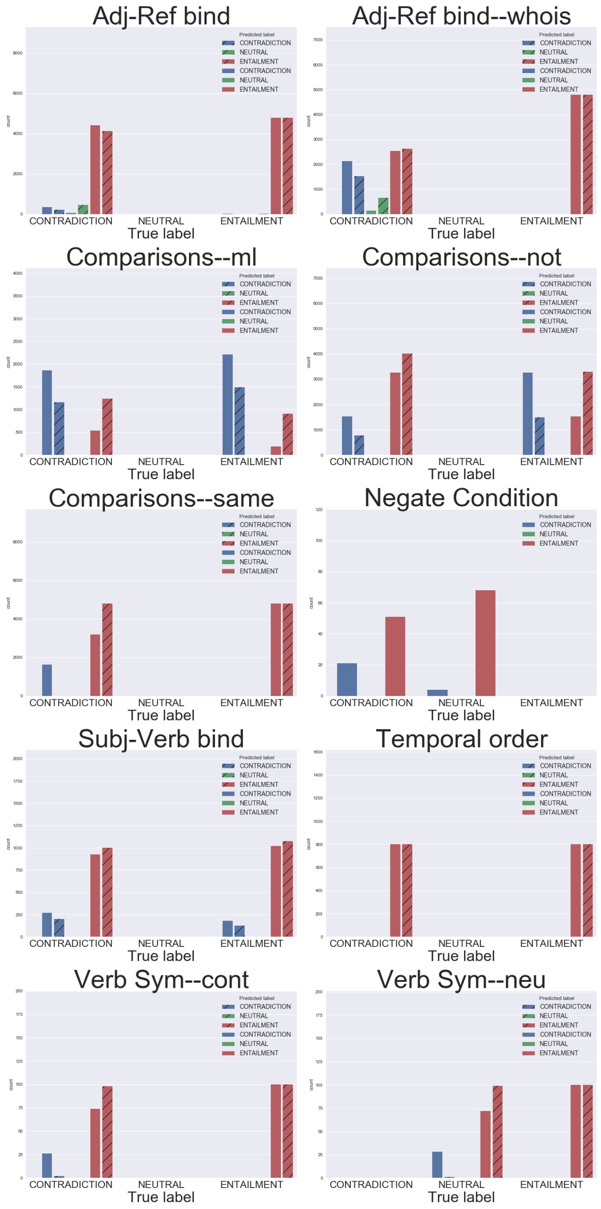

Figure 12: ScrambleTest Data: The classifications made by the logistic regression on the InferSent vectors. Long noun phrases are hatched.