

What If We Simply Swap the Two Text Fragments? A Straightforward yet Effective Way to Test the Robustness of Methods to Confounding Signals in Nature Language Inference Tasks

Haohan Wang¹, Da Sun², Eric P. Xing^{1,3}

¹Language Technologies Institute, School of Computer Science, Carnegie Mellon University,
Pittsburgh, PA, USA

²School of Information Science, Southeast University
Nanjing, China

³Machine Learning Department, School of Computer Science, Carnegie Mellon University,
Pittsburgh, PA, USA
haohanw@cs.cmu.edu

Abstract

Nature language inference (NLI) task is a predictive task of determining the inference relationship of a pair of natural language sentences. With the increasing popularity of NLI, many state-of-the-art predictive models have been proposed with impressive performances. However, several works have noticed the statistical irregularities in the collected NLI data set that may result in an over-estimated performance of these models and proposed remedies. In this paper, we further investigate the statistical irregularities, what we refer as confounding factors, of the NLI data sets. With the belief that some NLI labels should preserve under swapping operations, we propose a simple yet effective way (swapping the two text fragments) of evaluating the NLI predictive models that naturally mitigate the observed problems. Further, we continue to train the predictive models with our swapping manner and propose to use the deviation of the model’s evaluation performances under different percentages of training text fragments to be swapped to describe the robustness of a predictive model. Our evaluation metrics leads to some interesting understandings of recent published NLI methods. Finally, we also apply the swapping operation on NLI models to see the effectiveness of this straightforward method in mitigating the confounding factor problems in training generic sentence embeddings for other NLP transfer tasks.

Natural Language Inference (NLI) task is testing the ability of a computational model to understand the natural language by the evaluation of a three-way classification for two fragments of natural language texts (Bowman et al. 2015). The two fragments, namely *premise* (P) and *hypothesis* (H), are labeled in three different categories (*i.e.* *entailment* (E), *neutral* (N), and *contradiction* (C)) based on their entailment relation. More specifically, we can have:

- *E*: *H* is definitely true given *P*
- *C*: *H* is definitely not true given *P*
- *N*: *H* may or may not be true given *P*

Formally, we could rewrite the natural language described linguistic entailment relation into propositional logic as following:

To appear at Association for the Advancement of Artificial Intelligence (AAAI) 2019.

- $E: P \rightarrow H$
- $C: P \rightarrow \neg H$
- $N: P \perp H$

where \rightarrow stands for implication, \neg stands for negation, and we use \perp to denote that there is no clear relation between *P* and *H*.

The essential part of this paper lies in the fact that for any two propositions *A* and *B*, we have:

- $(A \rightarrow B) \perp (B \rightarrow A)$
- $(A \rightarrow \neg B) \iff (B \rightarrow \neg A)$
- $(A \perp B) \iff (B \perp A)$

In simpler words, swapping *A* and *B* will retain the \rightarrow relation and \perp relation, but not \rightarrow relation. Therefore, we can simply evaluate an NLI predictive model by swapping the *premise* and *hypothesis* in testing data set with the argument: **If a model can truly predict inference relationship between pairs of text fragments, it should report comparable accuracy between the original test set and swapped test set for *contradiction* pairs and *neutral* pairs, and lower accuracy in swapped test set for *entailment* pairs.** If we do not observe such an accuracy pattern, it is very likely that the evaluated NLI predictive model is questionable despite its impressive performance. One may argue that the label may not necessarily preserve for *neutral* pairs after swapping, we explain in the Discussion section later.

This work is inspired by several recent works that have observed the statistical irregularities of constructed NLI data sets. For example, Gururangan et al. (2018) noticed that during the data construction phase, the workers create a heuristic to generate hypothesis with minimum risks of being wrong, which lead to the fact that there are different distributions of words for different labels, and a powerful model can easily bypass the semantic information of the texts and predict the label with reasonable accuracy. The similar phenomenon has been observed by Poliak et al. (2018), who noticed that a model could infer the NLI label with only hypotheses, and proposed a hypothesis-only baseline. Recently, Naik et al. (2018) also noticed that the machine learning models might exploit the idiosyncrasies in the construction of the data set in predicting NLI labels and proposed

a “stress test” to evaluate whether the models can make semantic-level inferences. All these previous work have discussed the inherent limitations of current NLI tasks and proposed solutions such as more comprehensive evaluation criteria (Naik et al. 2018), guidance in future constructions of data sets (Gururangan et al. 2018), or more powerful methods such as utilizing the propositional logic relationship to regularize the neural network to ignore statistical irregularities of words and predict through semantic-level information (Minervini and Riedel 2018). While this paper agrees with the necessity of these carefully-designed remedies, we want to bring the community’s attention to a straightforward, easily-implemented, yet meaningful solution: swapping the two text fragments.

Our contribution of this paper can be summarized in three-fold:

- By recognizing the spurious signals between sentences’ semantic and words’ distribution of *hypothesis* created through a confounding factor of word choices, we introduce a straightforward and effective way to break such dependency: swapping the two text fragments. We apply the swapping evaluation metric on several recent NLI models and notice that some methods seem to predict based on local statistical patterns, in contrast to semantic level understanding.
- Inspired by the swapping evaluation metric, we propose to train the NLI models with a sequence of different training data sets that are defined by the percentages of sentence pairs swapped in the training set. We examine the “deviation” of performances of the model trained by these data sets evaluated by more powerful metrics. We propose the “deviation” to be a good measure of the robustness of the models to confounding signals.
- We further investigate how the swapping training procedure help mitigate the confounding factor problem during NLI training by applying the trained sentence embeddings to other NLP transfer tasks.

The remainder of the paper is organized as follows. We first introduce the background of this paper, where we explain why swapping the two text fragments is necessary, rather than just a random trick we adopt, and we also explain the concept of confounding factors in the context of NLI. Then we introduce the swap evaluation and the results, which further disclose some mechanism of the evaluated models. Further, we continue to test the robustness of the models to confounding signals. Afterward, we test to see if the swapping training in NLI will lead to more meaningful sentence embeddings that help in other NLP transfer tasks. Finally, with a brief discussion of the related work, we conclude the paper.

Background: Artifacts in NLI and Confounding Factors

The NLI data set (*e.g.* the SNLI data set introduced by Bowman et al. (2015)) is constructed with the help of Amazon Mechanical Turk workers by presenting the worker with one

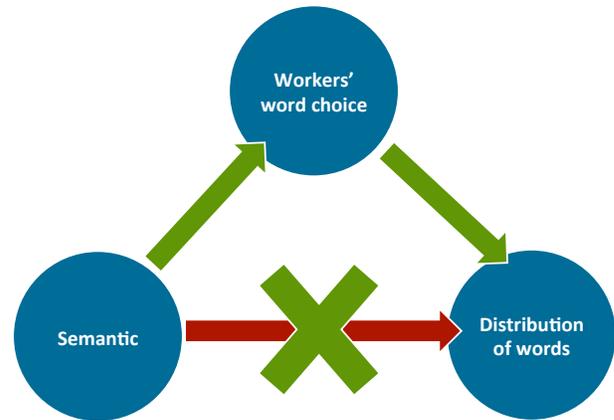


Figure 1: Illustration of the confounded relationship between the words’ distribution and the semantic information: The Amazon Mechanical Turk workers choose which words to use according to the inference relationship they are asked, and their choice affects the distribution of the words. As a result, a spurious relationship is created between the semantic and words’ distribution (denoted as the red arrow).

sentence (*premise*) and asking the workers to write sentences (*hypothesis*) that are:

- definitely true given the presented sentence, for *entailment*,
- definitely not true given the presented sentence, for *contradiction*
- might be true given the presented sentence, for *neutral*

As Gururangan et al. (2018) pointed out, these workers seem to find convenient ways to write *hypothesis*, such as using negation words (such as “no”, “nobody”, and “nothing”) to highlight the *contradiction* relation, or using generic words to replace specific words (such as “animal” for “dog”, “instrument” for “guitar”, and “outdoors” for “beach”) to guarantee the *entailment* relation. Therefore, these workers create different word distributions according to the different labels.

As a result, there exist confounding factors in the NLI data set. As we can see in Figure 1, the semantic label leads the workers’ choices of words, which further directly determines the word distributions of *hypothesis*. Therefore, it creates a spurious signal that the distribution of words, in addition to the semantic meaning of these sentences, is also related to the NLI label. If the machine learning models capture this spurious relation, the models will report impressive performances if evaluated regularly with NLI testing data set but will exhibit less favorable performances if evaluated with sophisticated methods.

By acknowledging these background information, one should notice that the swapping operation we propose in this paper is more than a simple data augmentation trick. It serves as an immediate solution for the aforementioned con-

founding factor problem: by swapping *premise* and *hypothesis* in test data set, we change the word distributions between training set *hypothesis* and test set *hypothesis*, therefore, models that can only predict through these spurious signals will unlikely be effective in the new test set.

Swapping Evaluation and Results

We proceed to officially introduce the Swapping evaluation method, what we expect, the results we have tested on recent published state-of-the-art methods, and follow-up analyses.

Despite the importance of correcting the artifacts, our method is as simple as swapping the *premise* and *hypothesis*. However, one should notice we do not always expect high scores in the evaluation. For a model that will predict the pairs on semantic levels, we expect:

- (significant) drop of performance for *entailment* pairs,
- roughly the same performance for *contradiction* pairs,
- roughly the same performance for *neutral* pairs.

Also, it's worth pointing out that a model that does not introduce performance drop in *entailment* pairs may not be used as an evidence to show the robustness of this model, but should be used as evidence to show that the model or the data may have some other interesting properties that we need further investigate.

We applied this evaluation method onto the following six different methods:

- CBOW: An MLP that uses averaged continuous bag of words as features.
- InferSent¹ (Conneau et al. 2017): It consists of sentence embedding, sequence encoder, composition layer, and the top layer classifier. The top layer classifier is an MLP whose input is a concatenation of *hypothesis* representation, *premise* representation, the dot product of these two, and the absolute value of the difference of these two.
- DGA (Deep Gated Attention Bidirectional LSTM)² (Chen et al. 2017): The structure is similar to InferSent. The composition layers involves an operation named Gated-attention, which is inspired by the fact that human tends to remember only parts of the sentence after reading.
- ESIM (Enhanced Sequential Inference Model)³ (Chen et al. 2016): A method that introduces local inference modeling, which models the inference relationship between *premise* and *hypothesis* after the two fragments aligned locally.
- KIM (Knowledge-based Inference Model)⁴ (Chen et al. 2018): This model enriches ESIM with external knowledge. At this moment, the external knowledge includes lexical semantic relation, including whether two words are synonymy, antonymy, hypernymy, hyponymy *etc.*

¹<https://github.com/facebookresearch/InferSent>

²https://github.com/lukec1231/enc_nli/

³<https://github.com/lukec1231/nli>

⁴<https://github.com/lukec1231/kim>

- ADV (Adversarially Regularized Neural NLI Model)⁵ (Minervini and Riedel 2018): an NLI model that is trained to minimize the standard loss as well as an inconsistency loss, which measures the model's performance over the adversarial set of data generated following logical rules.

The results⁶ are reported in Table 1. We can observe the performance drop in many cases, some of which are significant. Therefore, following our previous argument, the results, unfortunately, indicate the potential incompetence of some of these evaluated models in capturing the semantic-level information despite these models' impressive performances with the original evaluation metric.

It is relieving to notice that for *contradiction* pairs and *neutral* pairs, the models can still predict with an accuracy that is better than random chances. We believe this is because that the model can still capture a certain amount of semantic information from data to achieve an above-random prediction accuracy.

We notice that the performance drop of DGA is quite minor, which may indicate that the gated-attention mechanism can help exclude the signals from superficial statistical signals such as word distributions.

Also, we notice that for the methods InferSent and ESIM, the performance drop of *contradiction* pairs is greater than the performance drop of *neutral* pairs, but this trend is reversed for KIM and ADV. KIM and ADV are relatively new methods, therefore, generally believed to be more powerful than other methods. Interestingly, KIM relies on external knowledge in lexical semantic relation, and ADV is regularized by extra adversarial data that are generated according to logical rules. There might exist two explanations for KIM and ADV: 1) We conjecture that the extra information used by KIM and ADV helps eliminate the statistical irregularities in *contradiction* pairs, but introduce extra spurious signals for *neutral* pairs; 2) it is possible that KIM and ADV can more reliably ignore spurious signals than other methods because, as we will explain in detail in the Discussion section, some *neutral* pairs may not preserve the label after the swapping operation so certain amount of performance drop is expected.

Interestingly, we notice that there is barely any performance drop for ADV in the *entailment* pairs after swapping while all the other methods see a significant performance drop. As we argued previously, it might not be good if the performance does not drop when we expect so. This result indicates that we may need to investigate further into the mechanism of ADV.

While more error analysis are presented in the Appendix. To conclude this section, we propose a straightforward swapping evaluation method for methods of the NLI task. With our evaluation method, we can quickly tell some interesting properties of the current NLI methods.

⁵<https://github.com/uclmr/adversarial-nli/blob/master/nli/parser.py>

⁶All the experiments are run with standard Amazon EC2 p2.xlarge servers with Deep Learning AMI Version 13.0. The experiments are run with the default parameter settings set in the corresponding GitHub repositories.

Table 1: Swapping evaluation results on several NLI models. Swap- \star denotes the swapped data sets. Diff- \star denotes the performance drop of evaluated accuracy from the original data set to the swapped data set. We can observe the performance drop in many cases, some of which are significant. These results may indicate that these methods have been over-estimated in the ability in predicting the relationship of sentence pairs at the semantic level.

Model	Label	Dev	Swap-Dev	Diff-Dev	Test	Swap-Test	Diff-Test
CBOW	E	0.877	0.134	0.743	0.856	0.080	0.776
	C	0.706	0.583	0.123	0.740	0.580	0.160
	N	0.874	0.613	0.261	0.659	0.589	0.070
InferSent	E	0.850	0.090	0.760	0.880	0.087	0.793
	C	0.853	0.666	0.187	0.859	0.682	0.177
	N	0.795	0.713	0.082	0.795	0.712	0.083
DGA	E	0.822	0.376	0.446	0.854	0.422	0.432
	C	0.720	0.660	0.060	0.711	0.650	0.061
	N	0.700	0.648	0.052	0.700	0.619	0.081
ESIM	E	0.891	0.301	0.590	0.884	0.324	0.560
	C	0.865	0.702	0.163	0.861	0.701	0.160
	N	0.806	0.721	0.085	0.801	0.720	0.081
KIM	E	0.908	0.103	0.805	0.895	0.095	0.800
	C	0.850	0.772	0.078	0.845	0.796	0.049
	N	0.800	0.664	0.136	0.781	0.675	0.106
ADV	E	0.862	0.856	0.006	0.854	0.860	-0.006
	C	0.753	0.643	0.110	0.751	0.646	0.105
	N	0.706	0.509	0.197	0.705	0.507	0.198

Robustness Test to Confounding Factors Through Stress Test

Following previous sections where we showed that swapping evaluation can serve as an effective measure in determining whether a model predicts via spurious signals because it can break the dependency between sentence semantic and the words’ distribution of *hypothesis*, we continue to ask the question of whether the NLI models are robust to the confounding signals. The essential argument of this section is: **For an NLI model M , given a powerful metric T that can evaluate the NLI methods at the semantic level ignoring signals of confounding factors, if T reports similar performances of M_i , when M is trained repeatedly with different training data sets defined by the different percentages (i) of text pairs swapped, we can conclude that this model is robust to the confounding signals of words’ distribution.**

To explain this argument, if an NLI model M is learning through confounding signals, when such confounding signals are mitigated by swapping i percentage of the training pairs during training, we should observe changes of the evaluated performance when M_i is evaluated by T in comparison of M_0 evaluated by T . We define the overall magnitude of relative changes of M_i for multiple i as the robustness of M to the confounding signals of word distribution.

Also, note that a model is robust to confounding factors does not necessarily mean that the model can always predict via semantics. It only means the model is not focusing on spurious signals. It is possible that a model learns neither the semantic information nor the confounding information (e.g. a model that always predicts randomly).

Additionally, note that due to the label preserving prop-

erty for different labels we discussed in the previous section, only the *contradiction* and *neutral* pairs can be swapped during training.

To test the robustness of an NLI model, we need an evaluation metric that evaluates the model’s performance at the semantic level, independent of words’ distributions. We consider the recently proposed stress test (Naik et al. 2018). It is an evaluation method that helps to examine whether the models can predict at the semantic level. They created a test set that is constructed following a variety of different rules, including competence test set (antonym, numerical reasoning), distraction test set (with three strategies: word overlap, negation, and length mismatch), and also noise test set. There are six evaluation criteria altogether.

Following (Naik et al. 2018), we trained the NLI models discussed in the previous section on MultiNLI data set (Williams, Nangia, and Bowman 2017) and tested on the genre-matched and mismatched cases separately. We trained the models with five different cases (i.e. when 0%/25%/50%/75%/100% of training pairs are swapped) and reported the testing scores in Table 2. $S\star$ reports scores with different training data, and $R\star$ reports the relative changes of the corresponding case (ratio of $S\star$ over $S0\%$). We also calculated the deviation for each stress test, which indicates the robustness of the model to confounding factors (the smaller, the better). The deviation is calculated as the sum of the squared error between $R\star$ and 100% and is reported in the last column.

By looking into the last column of Table 2, we can see that different methods showed a different level of robustness. Overall, DGA is least affected by the spurious signals, which is consistent with our results in the previous section. For other methods, InferSent and ADV show robustness ex-

Table 2: ‘‘Stress Test’’ results for different percentage of training text fragments to be swapped. S^* denotes the percentage of text fragments are randomly swapped during training (Only *neutral* and *Contrdiction* pairs can be swapped). Accuracies shown on both genre-matched and mismatched categories for each stress test. R^* denotes the ratio of results over the case when no pairs are swapped ($S0\%$). *Devi* (last column) denotes the overall deviation of R^* from 100% within each test.

Model	Test	Cat	S0%	S25%	R25%	S50%	R50%	S75%	R75%	S100%	R100%	Devi
InferSent	Antonymy	Mat	22.87	24.92	109%	26.20	115%	27.61	121%	24.47	107%	0.173
		Mis	16.92	18.45	109%	19.61	116%	21.57	127%	17.82	105%	
	Length Mismatch	Mat	56.95	58.08	102%	57.11	100%	57.91	102%	57.26	101%	0.000
		Mis	58.31	58.84	101%	58.07	100%	58.94	101%	58.54	100%	
	Negation	Mat	48.78	49.60	102%	49.65	102%	49.45	101%	50.24	103%	0.002
		Mis	48.09	49.08	102%	49.15	102%	48.40	101%	48.99	102%	
	Word Overlap	Mat	55.58	56.86	102%	56.96	102%	54.21	98%	56.60	102%	0.004
		Mis	55.00	55.44	101%	56.05	102%	52.38	95%	55.51	101%	
Spelling Error	Mat	57.55	55.41	96%	55.80	97%	55.20	96%	55.34	96%	0.004	
	Mis	55.53	56.01	101%	55.31	100%	55.36	100%	54.95	99%		
DGA	Antonymy	Mat	15.02	14.32	95%	15.67	104%	14.02	93%	13.22	88%	0.073
		Mis	16.12	14.45	90%	16.38	102%	13.98	87%	13.13	81%	
	Length Mismatch	Mat	45.21	45.36	100%	44.27	98%	43.65	97%	42.19	93%	0.009
		Mis	43.97	45.21	103%	44.01	100%	42.16	96%	42.65	97%	
	Negation	Mat	44.91	43.28	96%	44.02	98%	43.99	98%	40.18	89%	0.025
		Mis	43.99	43.17	98%	43.87	100%	43.23	98%	39.09	89%	
	Word Overlap	Mat	52.39	50.56	97%	51.34	98%	52.42	100%	49.56	95%	0.005
		Mis	52.06	50.87	98%	51.65	99%	52.12	100%	50.23	96%	
Spelling Error	Mat	59.37	57.36	97%	58.11	98%	58.05	98%	56.36	95%	0.011	
	Mis	60.01	58.02	97%	58.92	98%	57.34	96%	55.67	93%		
ESIM	Antonymy	Mat	16.91	15.88	94%	16.07	95%	16.98	100%	13.23	78%	0.079
		Mis	16.7	15.29	92%	16.22	97%	15.87	95%	14.03	84%	
	Length Mismatch	Mat	44.98	40.66	90%	42.91	95%	45.43	101%	41.98	93%	0.016
		Mis	44.88	39.81	89%	41.99	94%	44.76	100%	41.73	93%	
	Negation	Mat	45.16	43.18	96%	44.67	99%	45.02	100%	43.01	95%	0.005
		Mis	45.27	43.29	96%	44.55	98%	45.00	99%	43.24	96%	
	Word Overlap	Mat	57.41	46.34	81%	44.35	77%	43.21	75%	37.73	66%	0.470
		Mis	58.27	46.20	79%	45.53	78%	43.40	74%	37.52	64%	
Spelling Error	Mat	57.09	50.66	89%	48.75	85%	43.65	76%	37.40	66%	0.370	
	Mis	56.48	50.98	90%	49.42	88%	44.31	78%	37.54	66%		
KIM	Antonymy	Mat	83.15	78.41	94%	76.69	92%	79.15	95%	76.49	92%	0.025
		Mis	83.04	78.43	94%	80.04	96%	81.08	98%	75.26	91%	
	Length Mismatch	Mat	47.27	46.89	99%	48.80	103%	48.23	102%	47.75	101%	0.003
		Mis	48.88	46.47	95%	48.90	100%	49.15	101%	46.96	96%	
	Negation	Mat	51.38	47.36	92%	40.91	80%	43.97	86%	38.86	76%	0.281
		Mis	53.04	48.18	91%	40.35	76%	45.51	86%	37.85	71%	
	Word Overlap	Mat	54.18	53.12	98%	54.32	100%	53.67	99%	52.87	98%	0.005
		Mis	54.57	53.27	98%	53.82	99%	52.55	96%	51.66	95%	
Spelling Error	Mat	60.19	59.16	98%	58.17	97%	59.98	100%	57.39	95%	0.010	
	Mis	61.32	60.36	98%	59.21	97%	59.65	97%	56.87	93%		
ADV	Antonymy	Mat	33.63	27.99	83%	25.82	77%	20.69	62%	21.46	64%	0.613
		Mis	30.22	29.76	98%	21.91	73%	20.88	69%	20.24	67%	
	Length Mismatch	Mat	36.23	35.98	99%	36.98	102%	34.12	94%	33.23	92%	0.022
		Mis	36.47	35.78	98%	37.04	102%	34.23	94%	33.45	92%	
	Negation	Mat	40.18	39.23	98%	37.38	93%	36.99	92%	34.19	85%	0.080
		Mis	40.86	39.48	97%	37.12	91%	36.78	90%	34.02	83%	
	Word Overlap	Mat	42.19	40.43	96%	39.67	94%	38.76	92%	36.43	86%	0.066
		Mis	41.98	40.43	96%	36.97	88%	38.76	92%	36.43	87%	
Spelling Error	Mat	37.89	35.17	93%	34.88	92%	34.12	90%	31.29	83%	0.097	
	Mis	37.63	35.26	94%	34.14	91%	34.09	91%	30.76	82%		

Table 3: Transfer test results for other downstream tasks when different percentage of training data are swapped.

	MR	CR	MPQA	SUBJ	TREC	SICK-E	SICK-R	MRPC
0%	76.92	78.15	87.64	90.79	83.6	82.3	0.859	74.09/ 82.45
25%	76.26	79.13	87.72	90.92	84.6	82.87	0.856	73.1/81.97
50%	77.3	79.84	87.5	90.28	79.8	83.17	0.862	75.25 /82.26
75%	76.42	79.66	87.71	90.99	80.4	83.24	0.859	73.8/82.32
100%	75.14	77.43	87.5	90.81	85	83.09	0.855	73.22/82.01

cept on Antonymy test, KIM shows robustness except on Negation test (which is also consistent with the findings in the previous section), and ESIM shows robustness except in Word Overlap and Spelling Error case.

We do not expect that swapping the training data pairs will help the model in dealing with spelling errors so that all the models should be robust in the Spelling Error test case. However, interestingly, ESIM shows a surprising drop of performance in Spelling Error test, and the magnitude of performance drop seems to correlate with the amount of data swapped. Other drops that are correlated with the amount of data swapped include the Word Overlap test for ESIM and Antonymy test for ADV.

Also, we notice KIM leads in test scores and the Antonymy test for KIM is impressively higher than any other competitors. This result is probably due to that KIM utilizes an external knowledge base that includes antonymy information.

Swapping Training for NLP Transfer Tasks

Finally, we study how the swapping operation will help mitigate the confounding factor problem by training a general-purpose sentence embedding that can capture the generic information for other transfer NLP tasks. Following (Conneau et al. 2017), we re-train the InferSent model with the swapping operation and compare the performance when different percentages of sentence pairs are swapped.

We evaluated the sentence embedding with a set of different transfer NLP tasks (Conneau et al. 2017), including several basic classification tasks such as sentiment analysis (MR, SST), question-type (TREC), product reviews (CR), subjectivity/objectivity (SUBJ) and opinion polarity (MPQA), paraphrase identification (MRPC), and entailment (SICK-E) and semantic relatedness (SICK-R) from SICK data set (Marelli et al. 2014).

These transfer tasks are evaluated with the standard metrics applied on these tasks. The MR, SST, TREC, CR, SUBJ and MPQA are evaluated with accuracy (Conneau et al. 2017). The MRPC is evaluated with both accuracy and F1 (Subramanian et al. 2018). The SICK-E and SICK-R are evaluated with Pearson correlation (Tai, Socher, and Manning 2015).

We report the results in Table 3. As we can see, although the improvement seems marginal, the swapping operation helps improve these transfer tasks because the model is often evaluated as the best when 25%-75% of the sentences are swapped.

Discussion

One may argue that the *neutral* pairs may not remain *neutral* after the two text fragments swapped, especially when the *hypothesis* is more specific than *premise*. For example, with a *premise* “I bought books”, several *hypothesis* can be generated to change the NLI label, such as “I bought 5 books”, “I bought books on history”, or “I bought books in the bookstore near campus”. However, through this paper, we assume the label preserves in most cases. This assumption is verified empirically by results in Table 1: the performance differences due to swapping are very similar in *contradiction* case and *neutral* case in most of the methods tested, leading to a conjecture that the label preserving property of swapping operation will be similar for *contradiction* and *neutral* in the data set.

Related Work

Computational Methods for NLI Ever since the introduction of recent large scale NLI data set (Bowman et al. 2015), many recent advanced computational models have been proposed, and a majority of these are LSTM methods or Bidirectional LSTM with extensions (Rocktäschel et al. 2015; Wang and Jiang 2015; Bowman et al. 2016; Liu et al. 2016a; Vendrov et al. 2015; Liu et al. 2016b; Liu, Qiu, and Huang 2016; Cheng, Dong, and Lapata 2016; Sha et al. 2016; Munkhdalai and Yu 2017; Munkhdalai and Yu 2017; Nie and Bansal 2017; Choi, Yoo, and Lee 2018; Peters et al. 2018).

Here we offer a brief overview of the most recent methods applied to NLI tasks: Tay, Tuan, and Hui (2017) proposed CAFE (Compare-propagate Alignment- Factorized Encoders). The essential component of CAFE is a compare-propagate architecture which first compares the two text fragments and then propagate the aligned features to upper layers for representation learning. Shen et al. (2018) presented reinforced self-attention (ReSA), which aims to combine the benefit of soft attention and a newly proposed hard attention mechanism called reinforced sequence sampling (RSS). They further plugged this ReSA onto a source2token self-attention model and applied to NLI tasks. Kim et al. (2018) proposed a densely-connected co-attentive recurrent neural network, whose essential idea is that the recurrent component uses the concatenated information from any layer to all the subsequent layers. They also used an autoencoder after dense concatenation to reduce the problem of ever-increasing sizes of representations. Chen, Ling, and Zhu (2018) introduced a vector-based multi-head attention as a generalized pooling method a weighted summation of hidden vectors to enhance sentence embedding. This

method is then built on a bidirectional LSTM and applied to NLI tasks. Tan et al. (2018) proposed a multiway attention network, which combines the information from four attention word-matching functions defined by four mathematical operations to build up the representation. They further built the proposed method on Gated Recurrent Unit (GRU) and applied to NLI tasks. Liu, Duh, and Gao (2018) introduced a stochastic answer network (SAN) for multi-step inference strategies for NLI. Different from conventional methods that directly predict given input sentence pairs, the SAN maintains a state and iteratively refines the predictions.

Confounding factor problems in other tasks and corresponding solutions In a broader domain other than NLI, many other machine learning methods noticed the problems introduced via confounding factors, which lead the machine learning methods yield a higher predictive performance than what a model can achieve in a confounder-free setting. For example, Wang et al. (2016) noticed that in video sentiment analysis, a random split of data into training set and testing set will typically yield much higher testing performance than split the data to make sure the data samples in testing set and training set will never come from the same individual (though the samples are different) because some recognizable features of the individual serve as confounding factors. Further, they proposed a select-additive learning method to mitigate the problem. Goyal et al. (2017) noticed that in visual question answering, the model would generate the answers for the images mainly based on the distribution of words of the training data, instead of the association between the images and the sentences in the training data. They further reorganize the data to balance the distributions to avoid such problems. The confounding-factor problem needs more attention in biomedical applications where usually all the side information such as gender and age can serve as confounding factors (Yue and Wang 2018). A recent paper empirically discusses the challenges (Zech et al. 2018), and a recent solution (Wang, Wu, and Xing 2018) mitigated the confounding factor challenge.

Conclusion

In this paper, we first discussed the existence of confounding factors of words' distribution of *hypothesis* in NLI data set due to the construction of data sets. Then with simple propositional logic rules, we presented a simple sanity check for the computational methods for natural language inference. Our argument is that for an evaluation data set where the *premise* and *hypothesis* are swapped, **if a model can truly predict inference relationship between pairs of text fragments, it should report comparable accuracy between the original test set and swapped test set for contradiction pairs and neural pairs, and lower accuracy in swapped test set for entailment pairs.** We applied our swapping evaluation towards several recently proposed models, and the results revealed some interesting properties of these methods.

Further, we proposed to train the NLI models with a sequence of different training data sets defined by the percentages of different sentence pairs swapped in the training set.

We then test these models with the stress test and investigate how the evaluated performance fluctuates for each model. We used "deviation" to measure the fluctuation, which is an indication of the robustness of the models to confounding factors of words' distribution. The robustness testing offered some other understandings of these NLI models. Overall, our swapping testing and robustness testing indicate that DGA (Chen et al. 2017) and KIM (Chen et al. 2018) are powerful at the semantic level and robust to the confounding factors. ADV (Minervini and Riedel 2018) is also a promising method, but some more detailed studies are recommended to be conducted for the interesting properties our metrics revealed.

Finally, we also tested to see how swapping operation can help mitigate the confounding factor problem by applying the trained sentence embedding to other NLP transfer tasks. We achieved a higher performance on most transfer tasks when 25%-75% of the sentence pairs are swapped.

References

- [Bowman et al. 2015] Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.
- [Bowman et al. 2016] Bowman, S. R.; Gauthier, J.; Rastogi, A.; Gupta, R.; Manning, C. D.; and Potts, C. 2016. A fast unified model for parsing and sentence understanding. *arXiv preprint arXiv:1603.06021*.
- [Chen et al. 2016] Chen, Q.; Zhu, X.; Ling, Z.; Wei, S.; and Jiang, H. 2016. Enhancing and combining sequential and tree lstm for natural language inference. *arXiv preprint arXiv:1609.06038*.
- [Chen et al. 2017] Chen, Q.; Zhu, X.; Ling, Z.-H.; Wei, S.; Jiang, H.; and Inkpen, D. 2017. Recurrent neural network-based sentence encoder with gated attention for natural language inference. *arXiv preprint arXiv:1708.01353*.
- [Chen et al. 2018] Chen, Q.; Zhu, X.; Ling, Z.-H.; Inkpen, D.; and Wei, S. 2018. Neural natural language inference models enhanced with external knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 2406–2417.
- [Chen, Ling, and Zhu 2018] Chen, Q.; Ling, Z.-H.; and Zhu, X. 2018. Enhancing sentence embedding with generalized pooling. *arXiv preprint arXiv:1806.09828*.
- [Cheng, Dong, and Lapata 2016] Cheng, J.; Dong, L.; and Lapata, M. 2016. Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*.
- [Choi, Yoo, and Lee 2018] Choi, J.; Yoo, K. M.; and Lee, S.-g. 2018. Learning to compose task-specific tree structures. In *Proceedings of the 2018 Association for the Advancement of Artificial Intelligence (AAAI), and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*.
- [Conneau et al. 2017] Conneau, A.; Kiela, D.; Schwenk, H.; Barrault, L.; and Bordes, A. 2017. Supervised learning of

- universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.
- [Goyal et al. 2017] Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, volume 1, 9.
- [Gururangan et al. 2018] Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S. R.; and Smith, N. A. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.
- [Kim et al. 2018] Kim, S.; Hong, J.-H.; Kang, I.; and Kwak, N. 2018. Semantic sentence matching with densely-connected recurrent and co-attentive information. *arXiv preprint arXiv:1805.11360*.
- [Liu et al. 2016a] Liu, P.; Qiu, X.; Chen, J.; and Huang, X. 2016a. Deep fusion lstms for text semantic matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 1034–1043.
- [Liu et al. 2016b] Liu, Y.; Sun, C.; Lin, L.; and Wang, X. 2016b. Learning natural language inference using bidirectional lstm model and inner-attention. *arXiv preprint arXiv:1605.09090*.
- [Liu, Duh, and Gao 2018] Liu, X.; Duh, K.; and Gao, J. 2018. Stochastic answer networks for natural language inference. *arXiv preprint arXiv:1804.07888*.
- [Liu, Qiu, and Huang 2016] Liu, P.; Qiu, X.; and Huang, X. 2016. Modelling interaction of sentence pair with coupled-lstms. *arXiv preprint arXiv:1605.05573*.
- [Marelli et al. 2014] 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*, 216–223.
- [Minervini and Riedel 2018] Minervini, P., and Riedel, S. 2018. Adversarially regularized neural nli models to integrate logical background knowledge. *arXiv preprint arXiv:1808.08609*.
- [Munkhdalai and Yu 2017] Munkhdalai, T., and Yu, H. 2017. Neural tree indexers for text understanding. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 1, 11. NIH Public Access.
- [Naik et al. 2018] Naik, A.; Ravichander, A.; Sadeh, N.; Rose, C.; and Neubig, G. 2018. Stress test evaluation for natural language inference. *arXiv preprint arXiv:1806.00692*.
- [Nie and Bansal 2017] Nie, Y., and Bansal, M. 2017. Shortcut-stacked sentence encoders for multi-domain inference. *arXiv preprint arXiv:1708.02312*.
- [Peters et al. 2018] Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.
- [Poliak et al. 2018] Poliak, A.; Naradowsky, J.; Haldar, A.; Rudinger, R.; and Van Durme, B. 2018. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*.
- [Rocktäschel et al. 2015] Rocktäschel, T.; Grefenstette, E.; Hermann, K. M.; Kočiský, T.; and Blunsom, P. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.
- [Sha et al. 2016] Sha, L.; Chang, B.; Sui, Z.; and Li, S. 2016. Reading and thinking: Re-read lstm unit for textual entailment recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2870–2879.
- [Shen et al. 2018] Shen, T.; Zhou, T.; Long, G.; Jiang, J.; Wang, S.; and Zhang, C. 2018. Reinforced self-attention network: a hybrid of hard and soft attention for sequence modeling. *arXiv preprint arXiv:1801.10296*.
- [Subramanian et al. 2018] Subramanian, S.; Trischler, A.; Bengio, Y.; and Pal, C. J. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. *arXiv preprint arXiv:1804.00079*.
- [Tai, Socher, and Manning 2015] Tai, K. S.; Socher, R.; and Manning, C. D. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.
- [Tan et al. 2018] Tan, C.; Wei, F.; Wang, W.; Lv, W.; and Zhou, M. 2018. Multiway attention networks for modeling sentence pairs. In *IJCAI*, 4411–4417.
- [Tay, Tuan, and Hui 2017] Tay, Y.; Tuan, L. A.; and Hui, S. C. 2017. A compare-propagate architecture with alignment factorization for natural language inference. *arXiv preprint arXiv:1801.00102*.
- [Vendrov et al. 2015] Vendrov, I.; Kiros, R.; Fidler, S.; and Urtasun, R. 2015. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*.
- [Wang and Jiang 2015] Wang, S., and Jiang, J. 2015. Learning natural language inference with lstm. *arXiv preprint arXiv:1512.08849*.
- [Wang et al. 2016] Wang, H.; Meghawati, A.; Morency, L.-P.; and Xing, E. P. 2016. Select-additive learning: Improving cross-individual generalization in multimodal sentiment analysis. *arXiv preprint arXiv:1609.05244*.
- [Wang, Wu, and Xing 2018] Wang, H.; Wu, Z.; and Xing, E. P. 2018. Removing confounding factors associated weights in deep neural networks improves the prediction accuracy for healthcare applications. *arXiv preprint arXiv:1803.07276*.
- [Williams, Nangia, and Bowman 2017] Williams, A.; Nangia, N.; and Bowman, S. R. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.
- [Yue and Wang 2018] Yue, T., and Wang, H. 2018. Deep learning for genomics: A concise overview. *arXiv preprint arXiv:1802.00810*.
- [Zech et al. 2018] Zech, J. R.; Badgeley, M. A.; Liu, M.; Costa, A. B.; Titano, J. J.; and Oermann, E. K. 2018. Confounding variables can degrade generalization performance of radiological deep learning models. *arXiv preprint arXiv:1807.00431*.

Appendix

Error Analysis of Table 1

While Table 1 has revealed several interesting properties of these NLI methods, an important question left is whether these drops of performance are because of the internal properties of the data sets, or more model-specific.

To answer this question, we collected the 400 sentence pairs that are most frequently mis-classified along all the epoches for each of those five models. We then investigate the overlaps of these samples and report the result in Table A1

Table A1: The overlaps of top 400 most frequently misclassification examples of each method once swapped

	InferSent	DGA	ESIM	KIM	ADV
InferSent	-	0	0	5	0
DGA	0	-	0	0	0
ESIM	0	0	-	0	5
KIM	5	0	0	-	7
ADV	0	0	5	7	-

As Table A1 shows, out of these 400 samples each, there are only 5 overlaps between InferSent and Kim, 5 overlaps between ADV and ESIM, and 7 overlaps between ADV and KIM. Therefore, the misclassification is more model specific.

We also calculate the distribution of six different types of misclassifications ($E \rightarrow N$, $E \rightarrow C$, $N \rightarrow E$, $N \rightarrow C$, $C \rightarrow E$, $C \rightarrow N$, where we denote the misclassification as “labels” \rightarrow “prediction”), and we notice that models showed a slight deviation from the averaged distribution overall. Particularly, ADV results in the least fraction in $C \rightarrow N$ out of these five methods, DGA results in the least fraction in $E \rightarrow N$ cases, and InferSent results in the least fraction in $N \rightarrow C$, as shown by Table A2.

Table A2: The fractions of mis-classification examples for each model

	InferSent	DGA	ESIM	KIM	ADV
$E \rightarrow N$	0.1175	0.0850	0.1125	0.1475	0.1075
$E \rightarrow C$	0.0325	0.0250	0.0425	0.0350	0.0275
$N \rightarrow E$	0.2050	0.2000	0.2250	0.2275	0.2225
$N \rightarrow C$	0.1800	0.2475	0.1825	0.2375	0.2050
$C \rightarrow E$	0.0325	0.0225	0.0375	0.0275	0.0225
$C \rightarrow N$	0.4325	0.4200	0.4000	0.3250	0.4150