Stress Test Evaluation for Natural Language Inference

Aakanksha Naik[1]*, **Abhilasha Ravichander**[1]*,
Norman Sadeh[2], **Carolyn Rose**[1], **Graham Neubig**[1]
[1]Language Technologies Institute, Carnegie Mellon University
[2]Institute of Software Research, Carnegie Mellon University
{anaik, aravicha, sadeh, cprose, gneubig}@cs.cmu.edu

Abstract

Natural language inference (NLI) is the task of determining if a natural language hypothesis can be inferred from a given premise in a justifiable manner. NLI was proposed as a benchmark task for natural language understanding. Existing models perform well at standard datasets for NLI, achieving impressive results across different genres of text. However, the extent to which these models understand the semantic content of sentences is unclear. In this work, we propose an evaluation methodology consisting of automatically constructed "stress tests" that allow us to examine whether systems have the ability to make real inferential decisions. Our evaluation of six sentence-encoder models on these stress tests reveals strengths and weaknesses of these models with respect to challenging linguistic phenomena, and suggests important directions for future work in this area.

1 Introduction

Natural language inference (NLI), also known as recognizing textual entailment (RTE), is concerned with determining the relationship between a premise sentence and an associated hypothesis. This requires a model to make the 3-way decision of whether a hypothesis is true given the premise (*entailment*), false given the premise (*contradiction*), or whether the truth value cannot be determined (*neutral*). NLI has been proposed as a benchmark task for natural language understanding research (Cooper et al., 1996; Dagan et al., 2006; Giampiccolo et al., 2007; Dagan et al., 2013; Bowman et al., 2015; Nangia et al.,

2017), due to the requirement for models to reason about several difficult linguistic phenomena[1] to perform well at this task (Bowman et al., 2015; Williams et al., 2017).

This benchmark task has been concretely implemented in the Stanford NLI (SNLI; Bowman et al. (2015)), and Multi-genre NLI (MultiNLI; Williams et al. (2017)) datasets, where crowdworkers are given a premise sentence and asked to generate novel sentences representing the three categories of entailment relations. Within the scope of these benchmark datasets, state-of-the-art deep learning-based sentence encoder models[2] (Nie and Bansal (2017), Chen et al. (2017), Conneau et al. (2017), Balazs et al. (2017), *inter alia*) have been shown to consistently achieve high accuracies, which may lead us to believe that these models excel at NLI across genres of text. However, machine learning models are known to exploit idiosyncrasies of how the data was produced, allowing them to imitate desired behavior using tricks such as pattern matching (Levesque, 2014; Rimell et al., 2009; Papernot et al., 2017). In NLI, this arises when the test set contains many easy examples, and the extent to which difficult components of language understanding are required is masked within traditional evaluation. Therefore we ask a natural question: is good model performance on NLI benchmarks a result of sophisticated pattern matching, or does it reflect true competence at natural language understanding?

In this work, we propose an evaluation based on "stress" tests for NLI, which tests robustness of NLI models to specific linguistic phenomena. This

---

[1]Such as scope, coreference, quantification, lexical ambiguity, modality and belief.

[2]Sentence-encoder models are considered to be especially important to natural language understanding due to the requirement to represent sentences as fixed length vectors and perform reasoning based on these representations (Nangia et al., 2017). Thus, they will be our primary focus.

*The indicated authors contributed equally to this work.

methodology is in part inspired by the work of Jia and Liang (2017) (and other related work: §5), which proposes the use of adversarial evaluation for reading comprehension by adding a distracting sentence at the end of a paragraph (known as concatenative adversaries), and evaluating models on this test set. However, this evaluation scheme cannot easily be applied to NLI as i) the adversarial perturbations suggested are not *label preserving*, that is the semantic relation between premise and hypothesis will not be maintained by such edits, ii) the perturbations may decrease system performance but are not interpretable and iii) premise-hypothesis pairs in inference usually consist of a single sentence, but concatenative adversaries break this assumption. In addition, besides evaluating model robustness to *distractions* in the form of adversarial examples, we are also interested in evaluating model *competence* on types of reasoning necessary to perform well at the task.

The proposed method offers a fine-grained evaluation for NLI in the form of "stress tests". Stress testing is a form of testing where systems are tested beyond normal operational capacity in order to identify weaknesses and to confirm that intended specifications are being met (Hartman and Owens, 1967; Tretmans, 1999; Beizer, 2003; Pressman, 2005; Nelson, 2009). To construct these datasets, we first examine the predictions of the best-performing sentence encoder model on MultiNLI (Nie and Bansal, 2017), and create a typology of phenomena that it finds challenging (§2). Based on this typology, we present a methodology to automatically construct stress tests, which may cause models that suffer from similar weaknesses to fail (§3). The resulting tests make it possible to perform evaluation on a phenomenon-by-phenomenon basis, which is not the case for Jia and Liang (2017).[3] We benchmark the performance of four state-of-the-art models on the MultiNLI dataset on our constructed stress tests (§4), and observe performance drops across stress tests. We view these results as a first step towards robust, fine-grained evaluation of NLI systems. To encourage development of models that perform true natural language understanding for NLI, we release our code and all stress tests for future evaluation.[4]

2 Weaknesses of State-of-the-art NLI

Before creating adversarial examples that challenge state-of-the-art systems, it is helpful to study what phenomena systems find difficult. To elucidate this, we conduct a comprehensive error analysis of the current best-performing sentence encoder model for MutliNLI (Nie and Bansal, 2017).

For this analysis, we sample 100 misclassified examples from both genre-matched and mismatched sets, analyze their potential sources of errors, and group them into a typology of common reasons for error. In the end, the reasons for errors can broadly be divided into the following categories[5], with examples shown in Table 1:

1. **Word Overlap (29%)**: Large word-overlap between premise and hypothesis sentences causes wrong entailment prediction, even if they are unrelated. Very little word overlap causes a prediction of neutral instead of entailment.

2. **Negation (13%)**: Strong negation words ("no", "not") cause the model to predict contradiction for neutral or entailed statements.

3. **Antonymy (5%)**: Premise-hypothesis pairs containing antonyms (instead of explicit negation) are not detected as contradiction by the model.

4. **Numerical Reasoning (4%)**: For some premise-hypothesis pairs, the model is unable to perform reasoning involving numbers or quantifiers for correct relation prediction.

5. **Length Mismatch (3%)**: The premise is much longer than the hypothesis and this extra information acts a distraction for the model.

6. **Grammaticality (3%)**: The premise or the hypothesis is ill-formed because of spelling errors or incorrect subject-verb agreement.

---

[3]Isabelle et al. (2017) have proposed a similar fine-grained evaluation approach for machine translation, but it requires the manual construction of examples, unlike our automatic approach.

[4]All stress tests can be found at goo.gl/vNLd7N . Code can be found at `https://github.com/AbhilashaRavichander/NLI_StressTest`

[5]% indicates what proportion of examples from our error analysis on the matched development set, fall into each category. Results on the mismatched development set follow similar trends.

Error	Premise	Hypothesis
Word Overlap (N→E)	And, could it not result in a decline in Postal Service volumes across–the–board?	There may not be a decline in Postal Service volumes across–the–board.
Negation (E→C)	Enthusiasm for Disney's Broadway production of The Lion King dwindles.	The broadway production of The Lion King is no longer enthusiastically attended.
Numerical Reasoning (C→E)	Deborah Pryce said Ohio Legal Services in Columbus will receive a $200,000 federal grant toward an online legal self-help center.	A $900,000 federal grant will be received by Missouri Legal Services, said Deborah Pryce.
Antonymy (C→E)	"Have her show it," said Thorn.	Thorn told her to hide it.
Length Mismatch (C→N)	So you know well a lot of the stuff you hear coming from South Africa now and from West Africa that's considered world music because it's not particularly using certain types of folk styles.	They rely too heavily on the types of folk styles.
Grammaticality (N→E)	So if there are something interesting or something worried, please give me a call at any time.	The person is open to take a call anytime.
Real World Knowledge (E→N)	It was still night.	The sun hadn't risen yet, for the moon was shining daringly in the sky.
Ambiguity (E→N)	Outside the cathedral you will find a statue of John Knox with Bible in hand.	John Knox was someone who read the Bible.
Unknown (E→C)	We're going to try something different this morning, said Jon.	Jon decided to try a new approach.

Table 1: Examples of misclassified samples for each error category (Gold Label→Predicted Label)

7. **Real-World Knowledge (12%)**: These examples are hard to classify without some real-world knowledge.

8. **Ambiguity (6%)**: For some instances, the correct answer is unclear to humans. These are the most difficult cases.

9. **Unknown (26%)**: No obvious source of error is discernible in these samples.

Some of these phenomena have long been suspected to be challenging for entailment models (Jijkoun and De Rijke, 2006; LoBue and Yates, 2011; Roy, 2017) while some others have been contemporaneously corroborated (Gururangan et al., 2018; Poliak et al., 2018). Phenomena such as negation scope and antonymy, are crucial for natural language understanding and have been of substantial interest in research in formal semantics (Kroch, 1974; Muehleisen, 1997; Murphy, 2003; Moscati, 2006; Brandtler, 2006).

3 Stress Test Set Construction

While the error analysis is informative for Nie and Bansal (2017), performing a manual analysis for every system is not scalable. To create an automatically calculable proxy, we focus on automatically constructing large-scale adversarial datasets (*stress tests*), which test NLI models on phenomena that account for most errors in our analysis. In particular, we generate adversarial examples

which test "word overlap", "negation", "length mismatch", "antonyms", "spelling error" and "numerical reasoning".[6]

We organize our stress tests into three classes, based on their perceived difficulty for the model. The first class (*competence tests*) evaluates the model's ability to reason about quantities and understand antonymy relations. The second class (*distraction tests*), estimates model robustness to shallow distractions such as lexical similarity or presence of negation words. This category contains "word overlap", "negation" and "length mismatch" tests. The final class (*noise tests*) checks model robustness to noisy data and consists of our "spelling error" test. Our adversarial construction uses three techniques: heuristic rules with external knowledge sources (for competence tests), a propositional logic framework (for distraction tests) and randomized perturbation (for noise tests). The following subsections describe our stress test construction, with examples shown in Table 2.

3.1 Competence Test Construction

Antonymy: For this construction, we consider every sentence from premise-hypothesis pairs in

---

[6]Notably, we focus only on "spelling error" for "grammaticality". We omit "real world knowledge" as it is not trivial to create a large dataset without human input, the "ambiguity" category because it is unreasonable that models can handle such cases, and the "unknown" category because it does not correspond to a particular phenomenon.

Error Cat.	Premise	Hypothesis
Antonyms	I love the Cinderella story.	I hate the Cinderella story.
Numerical Reasoning	Tim has 350 pounds of cement in 100, 50, and 25 pound bags	Tim has less than 750 pounds of cement in 100, 50, and 25 pound bags
Word Overlap	Possibly no other country has had such a turbulent history.	The country's history has been turbulent and true is true
Negation	Possibly no other country has had such a turbulent history.	The country's history has been turbulent and false is not true
Length Mismatch	Possibly no other country has had such a turbulent history and true is true and true is true and true is true and true is true and true is true	The country's history has been turbulent.
Spelling Errors	As he emerged, Boris remarked, glancing up at teh clock: "You are early	Boris had just arrived at the rendezvous when he appeared

Table 2: Example constructions from stress tests

the development set independently. We perform word-sense disambiguation for each adjective and noun in the sentence using the Lesk algorithm (Lesk, 1986). We then sample an antonym for the word from WordNet (Miller, 1995). The sentence with the word substituted by its antonym and the original sentence become a new premise-hypothesis pair in our set. This results in 1561 and 1734 premise-hypothesis pairs for matched and mismatched sets respectively.

Substituting a word with its antonym may not always result in a contradiction.[7] Hence, three annotators were provided 100 random samples from the stress test set to evaluate for correctness. At least two annotators agreed on 86% of the labels being contradiction. We also evaluated grammaticality of our constructions, with at least two annotators agreeing on 87% being grammatical.

Numerical Reasoning: Creating a stress test for numerical reasoning is non-trivial as most of the MultiNLI development set does not involve quantities. Hence, we extract premise sentences from AQuA-RAT (Ling et al., 2017a), a dataset specifically focused on algebraic word problems along with rationales for their solutions. However word problems from AQuA-RAT are quite complicated[8] and general-purpose NLI models cannot reasonably be expected to solve them.

To generate a reasonable set of premise sentences, we first discard problems which do not have numerical answers or have long rationales (>3 sentences) as such problems are inherently complex. We then split all problems into individual sentences and discard sentences without numbers, resulting in a set of 40,000 sentences. From this set, we discard sentences which do not contain at least one named entity (we consider "PERSON", "LOCATION" and "ORGANIZATION"), since such sentences mostly deal with abstract concepts.[9] This results in a set of 2500 premise sentences. For each premise, we generate entailed, contradictory and neutral hypotheses using heuristic rules:

1. **Entailment:** Randomly choose and change one numerical quantity from the premise, prefixing it with the phrase "less than" or "more than" based on whether the new number is higher or lower.

2. **Contradiction:** Perform one of two actions with equal probability: randomly choose a numerical quantity from the premise and change it, or randomly choose a numerical quantity from the premise and prefix it with "less than/ more than" without changing it.

3. **Neutral:** Flip the corresponding entailed premise-hypothesis pair.

Using these rules, we generate a set of 7,596 premise-hypothesis pairs testing models on their ability to perform numerical reasoning. We further instruct three human annotators to evaluate 100 randomly sampled examples for difficulty, grammaticality and label correctness (since the labels are automatically generated). At least two annotators agreed with our generated label for 91% of the samples. Additionally, at least two annotators

---

[7]Consider examples of sentences with modalities, belief, conjunction or even conversational text such as *"They can change the tone of people's voice yes."*, *"They can change the tone of people's voice no."*, coreference, word substitution in metaphors or failure of word sense disambiguation.

[8]Including concepts such as probability, geometry and theoretical proofs.

[9]For example, "Find the smallest number of five digits exactly divisible by 22, 33, 66 and 44".

agreed on 92% of the examples being grammatical, and 98% being trivial numerical reasoning for humans.

3.2 Distraction Test Construction

This class includes stress tests for "word overlap", "negation", and "length mismatch", which test model ability to avoid getting distracted by simple cues such as lexical similarity or strong negation words. Models usually exploit such cues to achieve high performance since they have strong but spurious correlations with gold labels, but reliance on shallow reasoning can be used to distract models easily, as we demonstrate. We use a novel framework inspired by propositional logic to construct adversarial examples.

Propositional Logic Framework: Assume a premise p and a hypothesis h. For entailment, $(p \Rightarrow h) \implies (p \wedge True \Rightarrow h)$ since $(p \wedge True = p)$. Similarly for the contradiction relation, $(p \Rightarrow\!\Leftarrow h) \implies (p \wedge True \Rightarrow\!\Leftarrow h)$ and if p and h are neutral, they remain neutral.

The next step is to construct statements whose values are true in all worlds (*tautologies*). We then define stress test accuracy for NLI as:

$$Adv(a) = \frac{1}{|D_{test}|} \sum_{i \in D_{test}} \mathbb{1}\{f(p, h, a) = c_i\}$$

where a is an adversarial tautology which can be attached as a conjunction to either the premise or hypothesis without changing the relation. We use this framework to construct example sets for "word overlap", "negation" and "length mismatch". For all sets, we use simple tautologies, which do not contain words that share any topical significance with the premise or hypothesis.

A natural concern is that statements obtained from such constructions are unnatural, making the NLI task more difficult for humans. To study this, we run a human evaluation where three annotators are shown premise-hypothesis pairs from these sets and instructed to label the relation. On word overlap, we find the provided label has 91% agreement with the gold label. For length mismatch, the provided label has 85% agreement with gold. This is similar to the agreement reported in (Williams et al., 2017), leading us to believe the constructed examples are not too unnatural or difficult. The constructions also remain grammatical; after annotating 100 samples from our adversar-

ially generated set only 2 were deemed ungrammatical, and both were because of reasons unrelated to our perturbations. Specific details for our sets are as follows:

Word Overlap: For this set, we append the tautology *"and true is true"* to the end of the hypothesis sentence for every example in the MultiNLI development set. Doing so reduces the word overlap between premise and hypothesis sentences, due to the introduction of additional words, in order to have an effect on models that rely on lexical overlap to make predictions.

Negation: For this set, we append the tautology (*"and false is not true"*), which contains a strong negation word (*"not"*), to the end of the hypothesis sentence for every example in the MultiNLI development set.

Length Mismatch: For this adversarial set, we append the tautology (*"and true is true"*) five times to the end of the premise sentence for every example in the MultiNLI development set. We modify the premise sentence in this case as we hypothesize that errors in this category mainly arise due to the premise sentence being unwieldy.

3.3 Noise Test Construction

This class consists of an adversarial example set which tests model robustness to spelling errors. Spelling errors occur often in MultiNLI data, due to involvement of Turkers and noisy source text (Ghaeini et al., 2018), which is problematic as some NLI systems rely heavily on word embeddings. Inspired by Belinkov and Bisk (2017), we construct a stress test for "spelling errors" by performing two types of perturbations on a word sampled randomly from the hypothesis: random swap of adjacent characters within the word (Example, *"I saw Tipper with him at teh movie."*), and random substitution of a single alphabetical character with the character next to it on the English keyboard. Example, *"Agencies have been further restricted and given less choice in selecting contractimg methods."*

4 Experiments

4.1 Experimental Setup

We focus on the following sentence-encoder models, which achieve strong performance on MultiNLI:

Nie and Bansal (2017) (NB): This model uses a sentence encoder consisting of stacked BiLSTM-

System	Original MultiNLI Dev		Competence Test			Distraction Test						Noise Test	
			Antonymy		Numerical	Word Overlap		Negation		Length Mismatch		Spelling Error	
	Mat	Mis	Mat	Mis	Reasoning	Mat	Mis	Mat	Mis	Mat	Mis	Mat	Mis
NB	74.2	74.8	15.1	19.3	21.2	47.2	47.1	39.5	40.0	48.2	47.3	51.1	49.8
CH	73.7	72.8	11.6	9.3	30.3	58.3	58.4	52.4	52.2	63.7	65.0	68.3	69.1
RC	71.3	71.6	36.4	32.8	30.2	53.7	54.4	49.5	50.4	48.6	49.6	66.6	67.0
IS	70.3	70.6	14.4	10.2	28.8	50.0	50.2	46.8	46.6	58.7	59.4	58.3	59.4
BiLSTM	70.2	70.8	13.2	9.8	31.3	57.0	58.5	51.4	51.9	49.7	51.2	65.0	65.1
CBOW	63.5	64.2	6.3	3.6	30.3	53.6	55.6	43.7	44.2	48.0	49.3	60.3	60.6
ALIGN	45.0	45.6	0.0	0.0	22.6	32.9	32.8	32.1	32.1	45.0	45.3	42.4	43.2

Table 3: Classification accuracy(%) of state-of-the-art models on our constructed stress tests. Accuracies shown on both genre-matched and mismatched categories for each stress set. For reference, random baseline accuracy is 33%.

RNNs with shortcut connections and fine-tuning of embeddings. It achieves the top non-ensemble result in the RepEval-2017 shared task (Nangia et al., 2017).

Chen et al. (2017) (CH): This model also uses a sentence encoder consisting of stacked BiLSTM-RNNs with shortcut connections. Additionally, it makes use of character-composition word embeddings learned via CNNs, intra-sentence gated attention and ensembling to achieve the best overall result in the RepEval-2017 shared task.

Balazs et al. (2017) (RiverCorners - RC): This model uses a single-layer BiLSTM with mean pooling and intra-sentence attention.

Conneau et al. (2017) (InferSent - IS): This model uses a single-layer BiLSTM-RNN with max-pooling. It is shown to learn robust universal sentence representations which transfer well across several inference tasks.

We also set up three simple baseline models:

BiLSTM: The simple BiLSTM baseline model described by Nangia et al. (2017).

CBOW: A bag-of-words sentence representation from word embeddings.

ALIGN: A bag-of-words alignment model inspired by the baseline proposed in MacCartney (2009). Our reimplementation uses Levenshtein edit distance for lexical similarity computation, but relies on POS tag information instead of IDF scores for term weighting. When trained on RTE3-dev, the model achieves an accuracy of 61.125% on RTE3 test (comparable to the reported average model performance in the RTE competition of 62.4%).

4.2 Model Performance on Stress Tests

Table 3 shows the classification accuracy of all six models on our stress tests and the original MultiNLI development set. We see that performance of all models drops across all stress tests. On competence stress tests, no model is a clear winner, with **RC** and **CH** performing best on antonymy and numerical reasoning respectively. On distraction tests, **CH** is the best-performing model. Interestingly, our **BiLSTM** baseline is the second-best model on two out of three distraction tests. On the noise test, **CH**, **RC** and both baselines **[BiLSTM;CBOW]** do not show much performance degradation, most likely due to the benefit of subword modeling via character-CNNs and the use of mean pooling. We further analyze model performance on each class of tests.

4.3 Model Competence

Model Performance on Antonymy: Table 3 shows that all models perform poorly on antonymy. In particular, **ALIGN** predicts the label "entailment" for all pairs due to high word alignment scores, resulting in 0% accuracy. **RC** achieves the best performance, with 36.4% and 32.8% on matched and mismatched sets respectively which is just higher than random performance. Our analysis shows that models tend to overpredict entailment (due to a high amount of word overlap in this test). This accounts for, on average, 86.4% and 87.6% of total errors on matched and mismatched sets.

We study which antonym pairs are easy and difficult for models by examining the errors of the best and worst performing neural models on this test **[RC;CH]**. On 982 samples where both mod-

els fail, we find 617 unique antonym pairs, and on 171 samples where both models succeed, we find 84 unique antonym pairs. 89.8% of the "easy" and 57.2% of the "hard" antonym pairs appear in a contradiction relation within the training data, suggesting that models succeed on easy antonym-pairs seen in the training data but struggle to generalize.

We were also curious about error variation by antonym type. We randomly sample 100 examples where both models fail and 100 samples where both succeed, and manually annotate whether the antonym present was gradable, relational or complementary. Among successful examples, 99% are complementary antonyms with only one relational antonym. Amongst the failure cases, 20% are relational antonym pairs, 73% are complementary and 7% are gradable, suggesting that models find relational and gradable antonyms hard, but get complementary antonyms both right and wrong. Finally, we examine differences between models by analyzing examples classified correctly by the best model which are not handled by the worst. We find that antonym pairs recognized by the weaker model occur, on average, nearly twice as often in the training data as antonym pairs recognized by the stronger model, suggesting that **RC** is able to learn antonymy from fewer examples (though these examples must be present in training data).

Model Performance on Numerical Reasoning: Table 3 shows that all models exhibit a significant performance drop on numerical reasoning, with none achieving an accuracy better than random (33%). We analyze the predictions of the best and worst performing models on this test **[BiLSTM;NB]**. The biggest source of common errors for both models (1703 out of 4337 errors) is misclassifying neutral pairs as entailment, which arises because our construction technique flips entailed premise-hypothesis pairs to create neutral pairs, leading to high word overlap for neutral pairs. Our constructions also lead to high word overlap for contradiction pairs, leading to a large number of C-E errors for both models (1695 out of 4337 errors). Most of the remaining errors are caused by entailment examples containing the phrases "more than" or "less than" being incorrectly classified as contradiction. This behavior could arise as these phrases are often used by

crowdworkers to create contradictory examples in the original MultiNLI data, fooling models into marking examples with this phrase as "contradiction" without reasoning about involved quantities. Our observations suggest that models do not perform quantitative reasoning, but rely on word overlap and other shallow lexical cues for prediction.

4.4 Model Distraction

Our distraction tests are designed to check model robustness to: 1) decreasing lexical similarity between premise-hypothesis pairs 2) strong negation words in sentence pairs.

Effect of Decreasing Lexical Similarity: Due to our construction methodology (appending tautologies), accuracy on *word overlap* and *length mismatch* demonstrates the effect of decreasing lexical similarity on model performance. Table 3, shows accuracy drops for all models on both tests. This drop is lower for **CH**, suggesting that their gated attention mechanism might help in focusing on relevant parts of the sentence. Moreover, **ALIGN** shows a much lower drop on *length mismatch* as compared to its performance drop on *word overlap*. This model gives a higher weight to $P(h|p)$ and hence shows higher performance drop when hypothesis sentences are altered (*word overlap*). The significant accuracy drop proves that lexical similarity is a strong signal for entailment prediction, failing which models default to neutral. To provide further justification, we compare the proportion of false neutral errors for all models on word overlap and length mismatch stress sets vs. the original MultiNLI set. As shown in table 4, we find it increases for all models, except **ALIGN** on both sets. For **ALIGN**, we only see an increase on *word overlap*

Effect of Introducing Strong Negation Words: Table 3 shows results on negation, and we see that all state-of-the-art models perform poorly, with accuracies dropping by 21.9% and 22.0%, on average, on matched and mismatched sets respectively. However, comparing the number of E-C (entailment predicted as contradiction) and N-C (neutral predicted as contradiction) errors for these models on the negation test vs. the original MultiNLI set, we do not find an increase in these error types on negation. Instead, we observe an

Sys tem	MultiNLI Dev		Word Overlap		Length Mismatch	
	Mat	Mis	Mat	Mis	Mat	Mis
NB	33.15	33.12	43.19	38.3	46.03	46.90
CH	32.92	31.65	84.66	85.33	65.81	65.75
RC	37.07	39.10	74.30	83.26	74.15	79.52
IS	34.73	31.42	86.26	87.04	43.48	44.16
BiLSTM	38.54	37.92	83.24	81.89	75.94	79.09
CBOW	33.86	30.20	74.46	72.25	54.69	59.93
ALIGN	67.25	64.10	95.20	95.90	56.90	59.70

Table 4: % of FALSE NEUTRAL in total errors on MultiNLI development set, word overlap test and length mismatch test.

increase in false neutral errors for all models. This could occur due to the introduction of extra words ("false", "is" and "true") apart from "not", indicating that decreasing lexical similarity has a stronger effect on models than introducing negation.

4.5 Effect of Noise

Our noise test results in Table 3 show that **NB** and **IS** exhibit a huge accuracy drop, since both models rely on word embeddings, and our perturbations often create words not present in standard english. Other models show little performance degradation on this test. **CH** performs subword modeling via character-level CNNs, which provides robustness towards perturbation attacks. **RC** and **BiLSTM** perform well despite relying on word embeddings since both use mean pooling, which might reduce the effect of single-word edits on the final representation. **CBOW** is also very robust to this test, which can arise from the fact that it sums word embeddings to create the final sentence embedding, diluting the effect of changing a single word on final model performance. **ALIGN** does not make use of word embeddings and relies on Levenshtein distance, providing excellent robustness to this particular test.[10]

4.6 Training with Distraction

Finally, we study the effect of training with distractions generated via our adversarial construction. We generate an equivalent sample with the negation distraction for every sample in the training data, and retrain **NB** and **BiLSTM** on the union of these examples and original training data.

---

[10]We analyze difference in model performance across perturbation techniques such as adjacent character swapping, keyboard character swapping, function word and content word perturbations, but do not observe significant changes.

System	BiLSTM		NB	
	Mat	Mis	Mat	Mis
MultiNLI Dev	70.14	70.36	66.60	66.63
NEGATION	68.91	70.43	49.31	48.68
DIFF TAUT	49.03	49.29	49.85	49.67

Table 5: Effect of training on distraction data on original DEV set, original distraction set and new distraction set

We observe the performance of the trained models on three tests: the original MultiNLI development set, the negation stress test and a new distraction test creating using a different tautology *"green is not red"* (DIFF TAUT) (Table 5). We observe that **NB** shows performance degradation across all tests, but training **BiLSTM** on distraction data helps it become robust to the tautology it was trained on. However, it collapses when evaluated on a different tautology. Ignoring such distractions is something humans do naturally. Models should not have to train on the specific distraction to succeed on this evaluation.

5 Related Work and Discussion

Adversarial evaluation schemes have been proposed to evaluate model robustness on various NLP tasks. Smith (2012) discuss dangers of community-wide "overfitting" to benchmark datasets and emphasize the need to correlate model errors to well-defined linguistic phenomena to understand specific model strengths and weaknesses. Prior work (Rimell et al., 2009; Schneider et al., 2017) performed analyses of model errors in dependency parsing and information extraction. Motivated by this desideratum, we analyze errors in Multi-Genre NLI and automatically construct large stress sets to evaluate NLI systems on identified difficulties. This is analogous to recent work (Jia and Liang, 2017; Burlot and Yvon, 2017) on developing automated adversarial evaluation schemes for reading comprehension and machine translation. However, unlike these efforts, our stress tests allow us to study model performance on a range of linguistic phenomena. Unlike work on manual construction of small adversarial evaluation sets for various NLP tasks (Levesque, 2014; Mahler et al., 2017; Staliūnaite and Bonfil, 2017; Isabelle et al., 2017; Belinkov and Bisk, 2017; Bawden et al., 2017), our work focuses on a more exhaustive large-scale evaluation for NLI. It

is interesting to note that Bayer et al. (2006) discuss the daunting cost of finding entailment pairs for NLI evaluation, but our techniques can be used to construct such pairs with low cost.

The NLI task attracted significant interest before datasets became large enough for the application of neural methods (Glickman et al., 2005; Harabagiu and Hickl, 2006; Romano et al., 2006; Dagan et al., 2006; Giampiccolo et al., 2007; Dagan et al., 2010; MacCartney, 2009; Zanzotto et al., 2006; Malakasiotis and Androutsopoulos, 2007; Haghighi et al., 2005; Angeli and Manning, 2014). de Marneffe et al. (2009) analyze the effect of multi-word expressions and find that they do not significantly affect the performance of NLI systems. Perhaps the closest to our contribution is the work of Marelli et al. (2014), which aims to construct sentences that require compositional knowledge (SICK). Our work extends this by specifically targeting identified weaknesses of neural models, and offering fine-grained evaluation on the same phenomena. The constructions also differ from SICK in two aspects. Since we use large datasets (Williams et al., 2017; Ling et al., 2017b) as base data, our sets are more lexically diverse than SICK which used a seed set of 1500 sentences. Secondly, while SICK uses handcrafted rules and incorporates linguistic phenomena, sentence-pairs are not constrained to exhibit only one phenomenon, which may introduce confounding factors during analysis. This observation also holds for the small portion of the MultiNLI development set annotated for particular linguistic phenomena. Conversely, our techniques force each evaluation set to focus on a single phenomenon, providing a testbed for fine-grained evaluation and analysis. Presence of biases in the MultiNLI development and test data (Gururangan et al., 2018; Poliak et al., 2018) could also lead to models exploiting them as shallow cues for prediction (for example, the performance of baseline models on the subset of the MultiNLI dataset annotated for antonymy averages 67% , while the same baselines perform much worse on our antonymy stress-test). The work of Cooper et al. (1996) on the FraCaS dataset identifies phenomena that a NLI system should be able to handle. However, though the dataset offers fine-grained evaluation, it requires examples to be manually constructed, making it difficult to construct large-scale suites of examples. In addition, our stress tests are meant to safeguard against identified pattern-matching behaviors of neural models, which is becoming increasingly relevant with the advent of deep-learning techniques for natural language inference. We hope that insights derived from our stress tests will stimulate future research in NLI. One promising direction would be the development and investigation of more linguistically-motivated neural models on NLI (such as models which incorporate explicit negation scope information or semantic roles for example), as our stress tests now provide a framework for in-depth analysis of model performance and demonstrate significant room for improvement in these areas. While we benchmark the performance of state-of-the-art models on our stress tests, in the future, it would be interesting to investigate which architectural choices contribute to model successes and *why*. Another interesting research direction is the identification of "core competencies", such as quantitative reasoning or antonymy, which can enhance model performance across multiple NLP tasks such as sentiment analysis, question answering and relation extraction and studying transfer of representations from "competent" models.

6 Conclusion

In this work, we present a suite of large-scale stress tests to perform targeted evaluation of NLI models, along with a set of techniques for their automatic construction. Our stress tests evaluate a model's ability to reason about quantities and antonymy (*competence tests*), its susceptibility to shallow lexical cues (*distraction tests*) and its robustness to random perturbations (*noise tests*). We benchmark the performance of four state-of-the-art sentence encoding models on our tests and find that they struggle on many phenomena, despite reporting high accuracy on NLI.

Overall, we consider the MultiNLI dataset to be a valuable resource for the NLP community, with entailment pairs drawn from several different genres of text. However, we argue that the community would benefit by having NLI models pass sanity checks, in the form of "stress tests", to ensure models evolve against exploiting simple idiosyncrasies of training data. Similar to Isabelle et al. (2017), we intend our stress tests to supplement existing NLI evaluation rather than replace it. In the future, we hope benchmarking model

performance on stress tests in addition to standard evaluation criteria will provide deeper insight into model strengths and weaknesses, and guide more informed model choices. We would also like to note that the "stress test" evaluation paradigm that we propose for NLI can be further updated in the future when new forms of models are devised, increasing the coverage of the tests to cover problems of future models as well. We release all our stress tests and associated code to the community to promote work on models which will get us closer to true natural language understanding.

Acknowledgements

References

Gabor Angeli and Christopher D Manning. 2014. Naturalli: Natural logic inference for common sense reasoning. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. pages 534–545.

Jorge Balazs, Edison Marrese-Taylor, Pablo Loyola, and Yutaka Matsuo. 2017. Refining raw sentence representations for textual entailment recognition via attention. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*. Association for Computational Linguistics, Copenhagen, Denmark, pages 51–55. http://www.aclweb.org/anthology/W17-5310.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2017. Evaluating discourse phenomena in neural machine translation. *arXiv preprint arXiv:1711.00513* .

Sam Bayer, John Burger, Lisa Ferro, John Henderson, Lynette Hirschman, and Alex Yeh. 2006. Evaluating semantic evaluations: How rte measures up. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, Springer, pages 309–331.

Boris Beizer. 2003. *Software testing techniques*. Dreamtech Press.

Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173* .

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 632–642. http://aclweb.org/anthology/D15-1075.

Johan Brandtler. 2006. On aristotle and baldness: Topic, reference, presupposition of existence, and negation. *Working papers in Scandinavian syntax* 77:177–204.

Franck Burlot and François Yvon. 2017. Evaluating the morphological competence of machine translation systems. In *Proceedings of the Second Conference on Machine Translation*. pages 43–55.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Recurrent neural network-based sentence encoder with gated attention for natural language inference. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*. Association for Computational Linguistics, Copenhagen, Denmark, pages 36–40. http://www.aclweb.org/anthology/W17-5307.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 670–680. https://www.aclweb.org/anthology/D17-1070.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. 1996. Using the framework. Technical report.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. The fourth pascal recognizing textual entailment challenge. *Journal of Natural Language Engineering* .

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, Springer, pages 177–190.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies* 6(4):1–220.

Marie-Catherine de Marneffe, Sebastian Padó, and Christopher D Manning. 2009. Multi-word expressions in textual inference: Much ado about nothing? In *Proceedings of the 2009 Workshop on Applied Textual Inference*. Association for Computational Linguistics, pages 1–9.

Reza Ghaeini, Sadid A Hasan, Vivek Datla, Joey Liu, Kathy Lee, Ashequl Qadir, Yuan Ling, Aaditya Prakash, Xiaoli Z Fern, and Oladimeji Farri. 2018. Dr-bilstm: Dependent reading bidirectional lstm for natural language inference. *arXiv preprint arXiv:1802.05577* .

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*. Association for Computational Linguistics, pages 1–9.

Oren Glickman, Ido Dagan, and Moshe Koppel. 2005. Web based probabilistic textual entailment .

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324* .

Aria Haghighi, Andrew Ng, and Christopher Manning. 2005. Robust textual inference via graph matching. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Sanda Harabagiu and Andrew Hickl. 2006. Methods for using textual entailment in open-domain question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 905–912.

Philip H. Hartman and David H. Owens. 1967. How to write software specifications. In *Proceedings of the November 14-16, 1967, Fall Joint Computer Conference*. ACM, New York, NY, USA, AFIPS '67 (Fall), pages 779–790. https://doi.org/10.1145/1465611.1465713.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 2486–2496. https://www.aclweb.org/anthology/D17-1263.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 2021–2031. https://www.aclweb.org/anthology/D17-1215.

Valentin Jijkoun and Maarten De Rijke. 2006. Recognizing textual entailment: Is word similarity enough? In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, Springer, pages 449–460.

Anthony S Kroch. 1974. *The semantics of scope in English.*. Ph.D. thesis, Massachusetts Institute of Technology.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*. ACM, New York, NY, USA, SIGDOC '86, pages 24–26. https://doi.org/10.1145/318723.318728.

Hector J Levesque. 2014. On our best behaviour. *Artificial Intelligence* 212:27–35.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017a. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 158–167. http://aclweb.org/anthology/P17-1015.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017b. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *arXiv preprint arXiv:1705.04146* .

Peter LoBue and Alexander Yates. 2011. Types of common-sense knowledge needed for recognizing textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, pages 329–334.

Bill MacCartney. 2009. *Natural language inference*. Stanford University.

Taylor Mahler, Willy Cheung, Micha Elsner, David King, Marie-Catherine de Marneffe, Cory Shain, Symon Stevens-Guille, and Michael White. 2017. Breaking nlp: Using morphosyntax, semantics, pragmatics and world knowledge to fool sentiment analysis systems. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*. Association for Computational Linguistics, Copenhagen, Denmark, pages 33–39. http://www.aclweb.org/anthology/W17-5405.

Prodromos Malakasiotis and Ion Androutsopoulos. 2007. Learning textual entailment using svms and string similarity measures. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. Association for Computational Linguistics, pages 42–47.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, Roberto Zamparelli, et al. 2014. A sick cure for the evaluation of compositional distributional semantic models.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

Vincenzo Moscati. 2006. *The scope of negation*. Ph.D. thesis, Università degli Studi di Siena.

Victoria Lynn Muehleisen. 1997. *Antonymy and semantic range in english*. na.

M Lynne Murphy. 2003. *Semantic relations and the lexicon: Antonymy, synonymy and other paradigms*. Cambridge University Press.

Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel Bowman. 2017. The repeval 2017 shared task: Multi-genre natural language inference with sentence representations. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*. Association for Computational Linguistics, Copenhagen, Denmark, pages 1–10. http://www.aclweb.org/anthology/W17-5301.

Wayne B Nelson. 2009. *Accelerated testing: statistical models, test plans, and data analysis*, volume 344. John Wiley & Sons.

Yixin Nie and Mohit Bansal. 2017. Shortcut-stacked sentence encoders for multi-domain inference. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*. Association for Computational Linguistics, Copenhagen, Denmark, pages 41–45. http://www.aclweb.org/anthology/W17-5308.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. 2017. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. ACM, pages 506–519.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042* .

Roger S Pressman. 2005. *Software engineering: a practitioner's approach*. Palgrave Macmillan.

Laura Rimell, Stephen Clark, and Mark Steedman. 2009. Unbounded dependency recovery for parser evaluation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*. Association for Computational Linguistics, pages 813–821.

Lorenza Romano, Milen Kouylekov, Idan Szpektor, Ido Dagan, and Alberto Lavelli. 2006. Investigating a generic paraphrase-based approach for relation extraction .

Subhro Roy. 2017. *Reasoning about quantities in natural language*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Rudolf Schneider, Tom Oberhauser, Tobias Klatt, Felix A. Gers, and Alexander Löser. 2017. Analysing errors of open information extraction systems. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*. Association for Computational Linguistics, Copenhagen, Denmark, pages 11–18. http://www.aclweb.org/anthology/W17-5402.

Noah A Smith. 2012. Adversarial evaluation for models of natural language. *arXiv preprint arXiv:1207.0245* .

Ieva Staliūnaite and Ben Bonfil. 2017. Breaking sentiment analysis of movie reviews. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*. pages 61–64.

Jan Tretmans. 1999. Testing concurrent systems: A formal approach. In *International Conference on Concurrency Theory*. Springer, pages 46–65.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426* .

F Zanzotto, Alessandro Moschitti, Marco Pennacchiotti, and M Pazienza. 2006. Learning textual entailment from examples. In *Second PASCAL recognizing textual entailment challenge*. PASCAL, page 50.