Evaluating adversarial attacks against multiple fact verification systems

James Thorne
University of Cambridge
jt719@cam.ac.uk

Andreas Vlachos
University of Cambridge
av308@cam.ac.uk

Christos Christodoulopoulos
Amazon
chrchrs@amazon.co.uk

Arpit Mittal
Amazon
mitarpit@amazon.co.uk

Abstract

Automated fact verification has been progressing owing to advancements in modeling and availability of large datasets. Due to the nature of the task, it is critical to understand the vulnerabilities of these systems against adversarial instances designed to make them predict incorrectly. We introduce two novel scoring metrics, attack *potency* and system *resilience* which take into account the correctness of the adversarial instances, an aspect often ignored in adversarial evaluations. We consider six fact verification systems from the recent Fact Extraction and VERification (FEVER) challenge: the four best-scoring ones and two baselines. We evaluate adversarial instances generated by a recently proposed state-of-the-art method, a paraphrasing method, and rule-based attacks devised for fact verification. We find that our rule-based attacks have higher potency, and that while the rankings among the top systems changed, they exhibited higher resilience than the baselines.

1 Introduction

Fact verification is the task of predicting whether claims can be supported or refuted by evidence. Advances in this task have been achieved through improved modelling and the availability of resources to train and validate systems (e.g. Wang (2017); Baly et al. (2018b); Thorne et al. (2018)). As this is a task with potentially sensitive applications like propaganda (Baly et al., 2018a) or biased news detection (Potthast et al., 2018), it is critical to understand how systems and models behave when exposed to real-world data and how deficiencies in their training data may contribute to this. It has been observed in related NLP tasks that as models become more complex, it is difficult to fully understand and characterize their behaviour (Samek et al., 2017). And from an NLP perspective, there has been an ongoing discussion as to what extent these models understand language (Jia and Liang, 2017) or they are exploiting unintentional biases and cues that are present in the datasets they are trained on (Poliak et al., 2018; Gururangan et al., 2018).

Original REFUTED Instance:
Bullitt is a movie directed by Phillip D'Antoni

Adversarial REFUTED Instance:
There is a movie directed by Phillip D'Antoni called Bullitt.

Adversarial SUPPORTED Instance:
Bullitt is not a movie directed by Phillip D'Antoni

Evidence:
Bullitt is a 1968 American action thriller film directed by Peter Yates and produced by Philip D'Antoni

Figure 1: Adversarial instances generated through rule-based transformations of existing claims

One of the diagnostic tools for understanding how models behave is *adversarial evaluation*, where data that is deliberately designed to induce classification errors is used to expose "blind spots" of a system. While there are many recently proposed techniques for generating adversarial instances for NLP tasks (surveyed in Section 2), they vary in the degree to which newly generated instances are correct, i.e. grammatical and appropriately labelled.

In this paper, we introduce two scoring metrics, (adversarial) attack *potency* and system *resilience*, that enable comparison of both adversarial instance generators and the systems that they are executed on respectively. We argue for manual evaluation of instances generated and the incorporation of their correctness for scoring. We conduct our experiments in the context of the FEVER Shared Task

(Thorne et al., 2018) (example in Figure 1), which incorporates both information retrieval and natural language inference to predict whether short factoid claims are SUPPORTED or REFUTED by evidence from Wikipedia. Systems must return not only the correct label but also the sentences providing the evidence for it. In the case where there is not enough evidence in Wikipedia for either label, the label NOTENOUGHINFO (NEI) is applied and no evidence needs to be returned.

We evaluate three adversarial attacks against four state-of-the-art systems and two baselines and apply our proposed scoring metrics that incorporate instance correctness. The first attack, informed by model behaviour, uses Semantically Equivalent Adversarial Rules (SEARs) (Ribeiro et al., 2018), a state-of-the-art method for generating rules that perform meaning-preserving transformations to instances that induce classification errors. The second attack is informed by dataset biases: we identify common patterns and constructions in the claims of the FEVER dataset and exploit them with hand-crafted rules to generate a number of new dataset instances. The final attack is a lexically-informed approach which makes use of a paraphrase model to generate new instances.

Our findings indicate that the instances generated by hand-crafted dataset-informed rules reduced all systems' classification accuracy more than the other approaches we tested. This was because the instances were not only more challenging to the systems under test, they were also more frequently correct – both of these characteristics factor into our *potency* score. While the lexically-informed approach induced a comparable number of misclassifications to the SEARs model, the instances were often incorrect, resulting in a *potency* score that was worse than sampling unmodified instances at random from the test set. Considering the *resilience* of the systems under adversarial evaluation, the 4^{th}-ranked model from the FEVER challenge performed better than the top system, but otherwise, the rankings of the systems were unchanged, and all top-ranked systems outperformed the baselines.

2 Methods for adversarial evaluation

Adversarial examples were initially studied in the field of computer vision (Szegedy et al., 2014) where model deficiencies such as over-sensitivity to perturbed inputs (by altering pixel intensities in a way which is imperceptible to humans) resulted in changes to the predictions. Making similar perturbations to text for adversarial evaluation is more challenging due to the discrete symbol space: modifying a single token may change an instance's semantics or make it ungrammatical. Various methods for attacks have been proposed ranging from manual construction and rule-based perturbation to automated paraphrasing and distractor information. Each of these methods can be informed by observations on the datasets for the task, model predictions and behaviour, and external knowledge and lexical resources. There is also a trade-off between the level of automation (which allows both scale and diversity of new claims) and to what extent the new instances are correct (i.e. grammatical and appropriately labelled). We survey a range of methods for generating adversarial instances and consider the suitability of the methods with respect to the FEVER task and generating correct instances.

Manual construction informed by expert knowledge: Instances that exploit world knowledge, semantics, pragmatics, morphology and syntactic variation have been written and compiled into challenge datasets for Machine Translation (Isabelle et al., 2017), Sentiment Analysis (Mahler et al., 2017; Staliūnaite and Bonfil, 2017) and Natural Language Understanding (Levesque, 2013). While these instances are expensive to construct, the attacker would have a high degree of confidence that the instances are correct and therefore correctness is not incorporated into the scoring metrics of any of these works. In FEVER, generating instances is more complex due to the need for annotators to highlight appropriate evidence: scaling up annotation to create new instances from scratch is non-trivial.

Noise introduced by character-level perturbations: Character-level attacks highlight the brittleness of systems by making letter swaps or insertions. Belinkov and Bisk (2018), Naik et al. (2018) and Ebrahimi et al. (2018) generate distorted examples which cause misclassifications or translation errors. An evaluation is performed comparing the performance of human crowdworkers for a classification task which only indicated minimal losses in classification accuracy between the original and modified instances. While it is unlikely that a single character can unintentionally change the semantics of a sentence, requiring relabelling, this method is still *intentionally* introducing typographical errors meaning that by the definition of the FEVER task, the instances would be incorrect.

Adding distractor information: Jia and Liang (2017) evaluated the addition of distractor sentences to instances and their effect on systems trained on the SQuAD question answering task (Rajpurkar et al., 2016). In their study, they compared appending a nonsensical string of words to the instance against appending sensible distractors generated by an automated method with the use of human annotators for a final filtering step. This approach cannot be ported to generate adversarial instances for the FEVER task as the claims are single sentences and the Wikipedia database which is used as evidence is considered immutable. Naik et al. (2018) concatenate manually-written logical tautologies (such as 'and true is true') to instances for a natural language inference task and this approach could be used to modify the claim portion of instances.

Paraphrasing existing instances: Iyyer et al. (2018) and Ribeiro et al. (2018) generate adversarial instances through paraphrasing existing ones. The method proposed by Ribeiro et al. (2018) is informed by both the predictions from a model as well as the external resource of a neural translation model whereas the approach taken by Iyyer et al. (2018) make use of a translation model only.

Zhao et al. (2018) use an autoencoder architecture to generate new instances that are semantically equivalent. Paraphrasing introduces the risk of meaning-altering changes to the sentence which would require relabelling, or producing sentences which are ungrammatical. Iyyer et al. (2018) identified that in at least 17.7% of cases, the generated examples were not paraphrases and in at least 14.0% of cases, the paraphrases were ungrammatical. In a pilot study, Zhao et al. (2018) find that 19% of generated examples were not 'semantically similar to the original input'[1] and ungrammatical in 14% of cases.

Ribeiro et al. (2018) generate textual replacement rules that are applied to dataset instances, modifying them in order to generate new semantically equivalent adversarial instances. In a crowdworker evaluation, not all rules are accepted as some induced higher high rates of incorrect instances: for the simpler task of sentiment analysis, 86.6% of rules were retained whereas, for visual question answering, only 43.5% of rules were retained. Paraphrasing approaches are applicable to FEVER, but their success depends on constructing instances that are both correctly labeled and grammatical. As FEVER consists of both information retrieval and natural language inference, the risk of generation of incorrect instances by paraphrasing can be quite high as subtle semantic changes may change the label of the instance.

3 Potency and resilience

Consider a method for generating adversarial instances (hereafter referred to as an adversary), a, that generates a set of instances $X_a = \{x_{a,i}\}_{i=1}^{N}$ with accompanying labels $Y_a = \{y_{a,i}\}_{i=1}^{N}$. To evaluate such adversaries, we must consider both their correctness and their effect on a system under test.

We measure the effectiveness of an adversary (a) through a system's (s) evaluation measure f (such as F_1 or FEVER Score), on a set of predictions made by the system $\hat{Y}_{s,a}$. Intuitively, better adversarial instances induce more misclassifications, resulting in a lower evaluation measure. Assuming the evaluation measure outputs a real value in the range $[0, 1]$, we define *potency* by the average reduction in score (from a perfect score) across all systems, $s \in S$ weighted by the correctness rate of the adversary, c_a.

$$\text{Potency}(a) \overset{\text{def}}{=} c_a \frac{1}{|S|} \sum_{s \in S} \left(1 - f(\hat{Y}_{s,a}, Y_a)\right) \quad (1)$$

Instances are correct if they are grammatical, appropriately labeled and meet the task requirements. To illustrate the need to incorporate correctness into evaluation, we will compare against *raw potency* where the correctness rate of the system c_a is set to 1. In our experiments, we will be using an estimate of the correctness rate for each adversary through annotating a sample of instances as it is not cost-effective with adversaries generating a large number of instances to perform a complete annotation.

A system that is resilient will have fewer errors induced by the adversarial instances, reflected in higher scores at evaluation. We wish to penalize systems for making mistakes on instances from adversaries with higher correctness rate. We define *resilience* as the average score scaled by the

[1] The authors do not state whether the newly generated adversarial instances retained the original label.

correctness rate for each adversary, $a \in A$:

$$\text{Resilience}(s) \overset{\text{def}}{=} \frac{\sum_{a \in A} c_a \times f(\hat{Y}_{s,a}, Y_a)}{\sum_{a \in A} c_a} \quad (2)$$

4 Adversarial attacks against fact verification systems

A FEVER dataset instance (example in Figure 1) consists of a claim accompanied by a label and evidence sentences unless the label is NOTENOUGHINFO (NEI) and no evidence needs to be provided. The adversaries we explore in this paper generate new instances by altering the claim sentence of existing instances and applying a new label if appropriate, without modifying the evidence. We chose this approach over generating completely novel instances as this obviates the need to search for new evidence, a process that is rather error-prone if done automatically. We explore three methods for generating adversarial instances by modifying existing dataset instances: manually crafted rule-based transformation informed by the training data; a recently proposed state-of-the-art method for automatically generating Semantically Equivalent Adversarial Rules (Ribeiro et al., 2018, SEARs) that is model targeted; and a lexically-informed method for paraphrasing model.

4.1 Training data-targeted adversary

This adversary assumes access to the dataset used to train the models and it identifies lexico-syntactic patterns in the claims that are highly frequent. The claims following these patterns undergo rule-based transformations to generate new instances with patterns not encountered in the training data. We evaluate three different types of transformations: entailment preserving rewrites, simple negations and complex negations (see examples in Table 1).

The entailment-preserving transformations we use include straightforward rules such as switching from active to passive voice while retaining the original label. For transformations that reverse the labels from SUPPORTED label to REFUTED and vice versa, negations were applied to a claim's verb phrase. More complex negations were introduced by combining the above two techniques.

We constructed 61 rules that matched the most common claim patterns identified in the training portion of the FEVER dataset, for example: X is a Y, X was directed by Y, X died on Y. 19 of our rules were entailment preserving, 19 were simple changes to negate claims and 23 were complex changes that also negated the claims. We release all of the rules on GitHub[2].

4.2 Lexically-informed adversary

Our adversary generates paraphrases of the claim without altering the evidence or label in a two-step process. The first step is to generate candidate claims x' by substituting nouns and adjectives with lemmas from all matching synsets from WordNet (Miller, 1995). The second step is to realize the surface forms for each of the lemmas. This is performed using a neural translation model, translating the candidate sentence into a foreign pivot language and back-translating into English yielding x'' using the model from Ribeiro et al. (2018). Instances are scored using the ratio of translation probabilities between the paraphrased and the original sentences $\frac{P(x''|x)}{P(x|x)}$ as given by the translation model. For instances where this ratio is low, the translated sentence is not representative of the semantics from the original claim. Where the ratio is high, we observed that the generated instances x'' were sometimes nonsensical: $P(x''|x)$ was extremely high in cases where sentences contained a high number of repeated words. To mitigate this, all claims were ranked using this measure and the upper and lower quartiles were excluded.

4.3 Model-targeting adversary

We also consider an adversary which exploits model predictions by generating Semantically Equivalent Adversarial Rules (Ribeiro et al., 2018, SEARs). The rules are generated by identifying transformations between paraphrased instances that alter the label predicted by a model while maintaining semantic equivalence. From all the rules generated from the dataset, a submodular selection is performed yielding a subset of non-redundant rules which induce misclassifications that apply to a large proportion of instances.

We applied three sets of adversarial rules to the FEVER dataset to generate new claims. The first set of adversarial rules induced classification errors in classifier for a sentiment analysis task (as reported in Ribeiro et al. (2018)) and the second and third sets of rules were generated using predictions made by the highest scoring model from the shared task (Nie et al., 2019).

[2] https://github.com/j6mes/fever-attacks-emlnp-2019

Transformation	Pattern	Template
Entailment Preserving	(.+) is a (.+)	There exists a $2 called $1
	(.+) (?:was | is)? directed by (.+)	$2 is the director of $1
Simple Negation	(.+) was an (.+)	$1 was not an $2
	(.+) was born in (.+)	$1 was never born
Complex Negation	(.+) (?:was | is)? directed by (.+)	There is a movie called $1 which wasn't directed by $2
	(.+) an American (.+)	$1 $2 that originated from outside the United States.

Table 1: Example rule-based attacks that preserve the entailment relation of the original claim (within the definition of the FEVER shared task), perform simple negation and more complex negations. The output new claims that are sufficiently different to confuse the classifier. The matching groups within the regular expression are copied into the template (variables begin with $).

Because SEARs requires $\mathcal{O}(n^2)$ model queries, generating rules using the full dataset may be prohibitively expensive for some users. In our evaluation, we compare the rules generated using model predictions on a sample of 1000 instances from the development set against rules generated from the full development set containing 9999 instances.

5 Experimental setup

We evaluate the potency of the adversarial instances generated by the approaches described in the previous section against four state-of-the-art models from the FEVER shared task and two baseline models. The models we evaluate were the Neuro-Semantic Matching Network (NSMN) (Nie et al., 2019), HexaF (Yoneda et al., 2018) Enhanced ESIM (Hanselowski et al., 2018) and the Transformer Model (Malon, 2018) as trained by the authors. For the baseline systems, we adopt the architecture from Thorne et al. (2018), using both a Decomposable Attention model and ESIM+ELMo model for natural language inference from AllenNLP (Gardner et al., 2017) and a Term Frequency - Inverse Document Frequency (TF-IDF) model for document retrieval and sentence selection.

Adversarial instances are generated by applying each adversary to existing FEVER instances and making modifications to the claim and label where appropriate. We apply each adversary to the development and test split of the dataset of Thorne et al. (2018)[3] to generate datasets of novel claims. We tune parameters using the development data and report results using a balanced uniform random sample of instances generated from the test data.

We measure instance correctness through blind annotation of a random sample of approximately 100 instances generated by each of the adversaries. We annotate each instance for grammaticality and whether the labelled claim is supported by the evidence from the original instance, following the manual error coding process described in Appendix B of the FEVER dataset description (Thorne et al., 2018).

6 Results

Applying each adversary to the 9999 instances in the FEVER test set generated a large number of new adversarial instances. From the training data targeted rule-based adversary, 21348 new claims were created, from the lexically-informed adversary using WordNet substitutions and a paraphrase model, 14046 instances were created, and using SEARs, 3668, 4144 and 7386 instances were generated depending on whether the rules were generated using the model of Nie et al. (2019) on the full dataset, sampled dataset or using the model of the sentiment analysis classifier respectively. For each adversary we create a balanced dataset for evaluation through a stratified sample of 1000 instances where we report the *potency* of each adversary and *resilience* of each system, considering instance correctness. Our summary results are presented in two tables: Table 2 describes the potency of each

[3] Note: This is the test split from the paper containing 9999 examples each that later formed part of a development set for the FEVER shared task. The reserved blind test instances from the shared task were not used.

Rank	Method	Raw Potency (%)	Correct Rate (%)	Potency (%)
1	Rule-based adversary	63.16	89.5	56.53
2	*Baseline (Unmodified Instances)*	44.79	97.0	43.46
3	SEARs (FEVER Full)	57.84	62.5	36.15
4	SEARs (FEVER Sample)	53.90	55.0	29.65
5	SEARs (Sentiment)	47.36	50.0	23.68
6	Paraphrase + WordNet	65.64	34.0	22.32

Table 2: Potency of adversaries where correctness rate is estimated using inspection of the generated instances. The baseline method of sampled instances is not used for scoring resilience. Raw potency is potency score without the considering the correctness of instances.

Rank	System	FEVER Score (%)	Resilience (%)
1	Transformer	57.04	58.66
2	NSMN	63.98	51.09
3	HexaF	62.34	50.06
4	Enhanced ESIM	61.32	43.98
5	TF-IDF + ESIM	36.87	26.86
6	TF-IDF + DA	27.71	22.28

Table 3: Systems ranked by the resilience to adversarial attacks. The FEVER Score column uses reported scores from the shared task.

adversary against the systems tested and Table 3 describes the resilience of each system to the adversaries.

Considering potency, the highest scoring adversary was the rule-based method – where the rules were manually written informed by observations on the training set. While the raw potency (not considering the correctness of instances) of the instances generated by the lexically-informed paraphrase adversary was higher than the rule-based adversary, a large number of instances generated from this method failed to meet the guidelines for the task, resulting in the lowest potency in our evaluation. We the adversarial instances that were generated by SEARs – which is informed by model behaviour – also had a higher raw potency than instances sampled from the dataset (baseline), the errors introduced when automatically generating adversarial instances resulted in a lower potency once the correctness of instances was taken into account. However, this method generated a higher rate of correct instances than the paraphrase based method.

For the rule-based method, the largest proportion of incorrect instances were generated by rules that inadvertently matched sentences that encountered in the observations used to design the rules. For example, the rule that transforms the pattern: 'X is a Y' to 'There is a Y called X' correctly generates instances in most cases, but for the sentence "Chinatown's producer is a Gemini" which contains a compound noun, the adversary "There is a Gemini called Chinatown's producer" was generated. Similar errors were produced when the patterns also matched determiners and included these in the new sentence in a position which altered the semantics or were ungrammatical.

The potency of SEARs is dependent on the model and dataset used for generating the Semantically Equivalent Adversarial Rules. When the rules were generated using the full FEVER development split, they not only resulted in a lower FEVER Score (contributing to higher potency) of the models but were also correct more often. Using only 10% of the FEVER development set instances when generating the adversarial rules was orders magnitude faster than the full dataset (generation took about a day instead of over two months), but this had a negative impact on the potency score in two ways: the systems made correct predictions on ap-

System	FEVER Score (%)				
	Rules	SEARS (FEVER Full)	SEARS (FEVER Sample)	SEARS (Sentiment)	Paraphrase
Transformer	56.36	58.26	62.66	68.67	44.24
NSMN	48.85	47.85	53.35	64.56	39.44
HexaF	45.25	50.15	55.06	62.86	35.64
Enhanced ESIM	31.53	46.75	48.05	62.26	38.24
TF-IDF + ESIM	20.72	28.13	31.03	31.03	27.83
TF-IDF + DA	18.32	21.82	26.43	26.43	20.72

Table 4: Breakdown of FEVER Scores of each system to each adversarial attack prior used for calculating resilience and potency. Lower scores indicates stronger attack (contributing to potency). Higher scores indicate stronger systems (contributing to resilience). Scores in this table do not account for correctness of instances.

proximately 4% more of these instances and the instances were less correct about 7% of the time. SEARs generated on the out-of-domain movie sentiment dataset had the lowest potency of the three variants we evaluated. Some of these rules (such as replacement of 'movie' with 'film') worked well when applied to FEVER.

A common failure mode from all variants of SEARs was the replacement of a large number of indefinite articles (such as *a*) with definite articles (such as *the*) making claims nonsensical or altering the semantics to the point where a label change was required. The SEARs also often made changes to determiners and quantifiers or deleted terms such as *only*, which altered the semantics of a claim, making it incorrectly labelled. A higher proportion of this behaviour was observed with SEARs generated from the sentiment analysis task.

The paraphrasing model exhibited a great degree of variance in the quality of the generated claims – resulting in a low correctness score. While all the claims generated were intelligible, there were a large number of grammatical errors present and some nonsensical paraphrases which altered the semantics by substituting synonyms for the wrong word senses. Some of the paraphrases resulted in proper nouns and named entities being replaced by third-person pronouns meaning that these claims failed to meet the FEVER guideline as all entities must be referenced directly to facilitate evidence retrieval.

The most resilient system was the transformer-based model (Malon, 2018) with a score of 58.66%. The instances generated by SEARs had a positive impact on the performance as both recall and FEVER Score were higher than the values reported from the shared task and the instances generated by manually constructed rules only negatively affected the model by a marginal amount. Meanwhile, the highest performing system from the shared task (NSMN) (Nie et al., 2019) had a lower resilience: 51.09%. Using the breakdown of FEVER Scores by system and adversary (reported in Table 4)), we observe that the NSMN was affected by the rule-based adversary and SEARS (FEVER Full) adversaries much more than transformer model.

With the exception of the transformer model, the resilience of the systems is well correlated with the performance on the original FEVER task. This demonstrates that the top-scoring systems in the shared task were able to better capture the semantic space of the task as the resilience of the baseline TF-IDF models and that evaluation was adequate in assessing their generalization abilities.

6.1 Effect of rule-based transformations

All the evaluated systems were pipelines of information retrieval and natural language inference components. We perform an analysis of the models under test considering the effect of rule-based adversarial instances and their impact on each pipeline stage. We compare the performance of the models using the new instances generated by applying transformations to existing instances (reported in *modified* columns) and compare this to the instances that were used to generate them (reported in *original* columns).

Considering the evidence retrieval component (refer to Table 5), we find that the adversarial instances did not affect the information retrieval component of some systems to the same extent as the NLI component. Most systems incorporated ei-

Model	Precision (%)			Recall (%)		
	Original	Modified	Delta	Original	Modified	Delta
Transformer	95.42	97.93	+2.51	71.47	50.00	-21.47
NSMN	41.14	44.18	+3.04	78.07	71.62	-6.45
HexaF	65.23	59.76	-5.47	80.78	75.98	-4.80
Enhanced ESIM	24.36	22.30	-2.06	86.04	80.93	-5.11
TF-IDF	10.49	9.10	-1.39	45.50	39.20	-6.30

Table 5: Effect of rule-based adversarial attacks on the evidence retrieval component of the pipelines considering sentence-level accuracy of the evidence.

Model	Accuracy (%)			FEVER Score (%)		
	Original	Modified	Delta	Original	Modified	Delta
Oracle + DA	82.38	62.56	-19.82	=	=	=
Oracle + ESIM	83.58	60.66	-22.92	=	=	=
TF-IDF + DA	46.49	37.44	-9.05	27.73	18.32	-9.41
TF-IDF + ESIM	49.75	37.34	-12.41	32.83	20.72	-12.11
Transformer	75.58	57.26	-18.32	73.87	56.36	-17.51
NSMN	70.27	50.35	-19.92	68.77	48.85	-19.92
HexaF	74.67	51.35	-23.32	67.37	45.25	-22.12
Enhanced ESIM	67.56	36.84	-30.72	62.76	31.53	-31.23

Table 6: Summary of label accuracy and FEVER Scores for instances used in rule-based adversarial attacks. For the case of the oracle evidence retrieval component, FEVER Score is equal to accuracy.

ther TF-IDF of keyword matching in their information retrieval component and thus they are little affected by the rule-based transformations (which were mostly adding stop words and reordering the words within the sentence). The only exception is the Transformer Model; even though it uses TF-IDF to retrieve documents, it makes use of an entailment classifier for sentence retrieval from the documents. While this approach maintained very high precision with the newly generated instances, the recall decreased indicating a brittleness for instances that differed from the distribution of the training set.

Considering the NLI stage of the systems (refer to Table 6), we observe that in an oracle environment (using manually labelled evidence – assuming perfect evidence retrieval), both the ESIM and Decomposable Attention models for natural language inference suffer a stark decrease in accuracy when making predictions on the adversarial examples. The decrease is less pronounced when considering the full pipeline (i.e. with an actual imperfect evidence retrieval component), which reduces the upper bound for the FEVER Score and accuracy. This is due to the noise introduced by the evidence retrieval pipeline stage.

The Enhanced ESIM system had the highest reduction in accuracy for the rule-based adversarial instances. Even though the model had an evidence recall of 80.93% (a reduction of 5.11%), the FEVER score reduced by 30.72%. On inspection, this model is mostly predicting SUPPORTED for most adversarial instances. In contrast, the Transformer Model had the lowest reduction in FEVER score (-17.51%) despite a reduction in evidence recall of 23.17%. This model also exhibited similar behaviour to the Enhanced ESIM model: while a large number of supported claims were correctly classified, this model predominantly predicted NOTENOUGHINFO for the adversarial instances. As this class does not require evidence to be correctly scored, ‘falling back’ to it resulted in a higher FEVER Score when no suitable evidence was found.

The rule-based adversary applies both entailment-preserving and label-altering (through negation) transformations to claims to generate new adversarial instances. We observe in all models (with the exception of the Enhanced ESIM (Hanselowski et al., 2018)) that the reduction in accuracy for the claims generated through entailment-preserving transformations was lower than the claims generated by the negating transformations. This may be revealing an inherent bias in the models similar to the one discussed by Naik et al. (2018), where models perform poorly for antonymous examples due to a dependence on word-overlap as a feature.

7 Conclusions

As automated means for generating adversarial instances do not always produce grammatical or correctly labelled instances, we developed a method for evaluation which incorporates instance correctness. We introduced two metrics that enable evaluation of adversarial attacks against systems: the *potency* of attack, and the *resilience* of systems to these attacks.

In this paper, we evaluated four state-of-the-art systems and two baselines against instances generated from three methods for adversarial attack: manually written rules that were informed by observations of the training data, an automated lexically-informed paraphrasing model and an automated state-of-the-art model-targeted approach. While the manually-written rules exhibited high correctness rate, the automated methods often made grammatical errors or label-altering changes which lowered the correctness rate. Our metrics captured this trade-off between correctness and impact on system scores – despite the paraphrasing automated method having the greatest impact on systems under test, it was not the most potent, owing to the low correctness rate.

We hope that the findings and metrics that we present in this paper help continue the discussion on improving the robustness of models and enable better comparisons between adversarial attacks to be made that consider instance correctness.

Acknowledgements

The authors wish to thank Trevor Cohn, Tim Baldwin and Oana Cocarascu for their helpful advice. The authors also wish to thank the teams UNC, UCLMR, Athene and Papelo of the FEVER1 shared task who made source code and pre-trained models available and were responsive to our questions.

References

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018a. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium. Association for Computational Linguistics.

Ramy Baly, Mitra Mohtarami, James Glass, Lluis Marquez, Alessandro Moschitti, and Preslav Nakov. 2018b. Integrating Stance Detection and Fact Checking in a Unified Corpus.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and Natural Noise Both Break Neural Machine Translation. In *ICLR*, pages 1–13.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-Box Adversarial Examples for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2017. AllenNLP: A Deep Semantic Natural Language Processing Platform.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation Artifacts in Natural Language Inference Data.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics.

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*, pages 2486–2496.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1875–1885.

Robin Jia and Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems.

Hector J Levesque. 2013. On our best behaviour. *IJCAI*.

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*, pages 33–39.

Christopher Malon. 2018. Team papelo: Transformer networks at fever. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 109–113, Brussels, Belgium. Association for Computational Linguistics.

George a. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress Test Evaluation for Natural Language Inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis Only Baselines in Natural Language Inference. (1):180–191.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. pages 2383–2392.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically Equivalent Adversarial Rules for Debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865. Association for Computational Linguistics.

Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *CoRR*, abs/1708.08296.

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.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *International Conference on Learning Representations*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426.

Takuma Yoneda, Jeff Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Ucl machine reading group: Four factor framework for fact finding (hexaf). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 97–102, Brussels, Belgium. Association for Computational Linguistics.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018. Generating Natural Adversarial Examples. *ICLR*.