Practical Attacks on Machine Translation Using Paraphrase

Elizabeth M. Merkhofer emerkhofer@mitre.org
John C. Henderson jhndrsn@mitre.org
Abigail S. Gertner gertner@mitre.org
Michael D. Doyle mdoyle@mitre.org
Lily Wong lwong@mitre.org
MITRE, McLean, Virginia, 22102, USA

Abstract

Studies show machine translation systems are vulnerable to adversarial attacks, where a small change to the input produces an undesirable change in system behavior. This work considers whether this vulnerability exists for attacks crafted with limited information about the target: without access to ground truth references or the particular MT system under attack. It also applies a higher threshold of success, taking into account both source language meaning preservation and target language meaning degradation. We propose an attack that generates edits to an input using a finite state transducer over lexical and phrasal paraphrases and selects one perturbation for meaning preservation and expected degradation of a target system. Attacks against eight state-of-the-art translation systems covering English-German, English-Czech and English-Chinese are evaluated under black-box and transfer scenarios, including cross-language and cross-system transfer. Results suggest that successful single-system attacks seldom transfer across models, especially when crafted without ground truth, but ensembles show promise for generalizing attacks.

1 Introduction

Recent studies show that natural language processing (NLP) applications are vulnerable to *adversarial perturbations*, where a small change to the input produces an undesirable change in system behavior, such as a lower-quality translation in a machine translation (MT) system (Ebrahimi et al., 2018; Cheng et al., 2019; Wallace et al., 2019; Cheng et al., 2020; Zhao et al., 2018; Zhang et al., 2021). These adversarial inputs offer insight into model robustness. They also can constitute vulnerabilities that expose everyday technology to malicious actors who would seek to deny and deceive artificial intelligence systems.

Practical concerns must be addressed to determine if these vulnerabilities persist outside of simplified scenarios. Most previous work uses the same ground truth to craft and evaluate an attack and relies on access to the model being attacked, such as model gradients (white-box) or the output of the model (black-box). We ask whether this vulnerability extends to attacks crafted with limited information about the target: without access to ground truth references, model weights or even the outputs of the particular MT system they are attacking. We examine transfer of adversarial examples among eight different MT systems with three target languages. For evaluation, we use a high threshold of success that takes into account both effect on translation quality and loss of meaning in the original text.

We introduce a novel text editing system (perturber) that rapidly generates hundreds or thousands of candidate edits using a compendium of vetted paraphrases scored to match human quality judgments. Adversarial edits are selected according to a configurable optimization trading off preservation of source-side meaning and degradation of target output. To simulate a scenario where a human reference is not available, the selector estimates degradation in translation quality using the change in translation output from a proxy MT system. These attacks meet the threshold for success when the MT system used for selection is matched to the victim model or when an ensemble of MT systems is used to do the targeting. However, we find that examples selected using a single translation model as proxy and ensembles crafted without sensitivity to source-side meaning changes do not often transfer to another victim model above the success threshold.

2 Practical Considerations

Overwhelmingly, previous work assumes high-information scenarios, using the same ground truth and model to craft and evaluate the attacks, and evaluates adversarial effect separately from the effect on the semantics of the input (Ebrahimi et al., 2018; Cheng et al., 2019; Wallace et al., 2019; Cheng et al., 2020; Zhao et al., 2018; Zhang et al., 2021). We address four considerations in evaluation of machine translation attacks with the purpose of understanding whether these attacks can be crafted in lower-information scenarios and whether the effect on system performance outweighs the degradation of the input text. First, we define our success criterion in a metric-independent way, drawing from Michel et al. (2019), to combine adversarial effect and degradation of the source in a single metric. Second, we calibrate similarity metrics so that one unit of meaning preservation in the source language side is as close as possible to one unit of translation quality in the target language side. Third, we consider whether attacks require access to ground truth to successfully degrade performance. Finally, we address whether attacks crafted using one system can be deployed against another to which it does not have access.

2.1 Successful MT Attacks

Effective adversaries do not simply change a system's behavior; they reliably degrade its performance. To attack MT, perturbations aim to maximally decrease translation quality with respect to the ground truth reference. The translations of a set of perturbed source segments should score lower than the originals under some MT metric, such as BLEU or CHRF. However, to ensure that the perturbations haven't trivially reduced translation quality by changing the meaning on the source side, we must also account for the effect of the perturbations on the meaning of the source.

We directly adopt several terms and metrics from Michel et al. (2019). We follow the convention that x and y refer to source and target language sentences, y_M is the translation produced by model M, and $\hat{x}$ and $\hat{y}_M$ are the perturbed version of the source sentence and its translation. We measure the translation quality of an attack by the *target similarity*, $s_{tgt}(y, \hat{y}_M)$, where y is a gold source reference translation and $\hat{y}_M$ is the MT system output on the perturbed input. The effect of a perturbation on the input text is measured by the *source similarity*, $s_{src}(x, \hat{x})$.

An attack degrades the target similarity by d_{tgt} in Equation 1. This is also referred to as *target relative score decrease*. It is similar to the version found in Michel et al. (2019) except we allow negative values of d_{tgt} if an attack inadvertently makes the translation *better*. We do this because we do not presume oracular access to a reference translation y at targeting time to decide when to avoid using a particular $\hat{x}$ for attack. A higher value of d_{tgt} means the output of the MT was more degraded. A value of zero means no degradation. Similarly, an attack degrades the source similarity by d_{src} in Equation 2. Using relative rather than absolute score decreases makes it possible to compare attack effectiveness across models with different original performance.

$$d_{tgt}(y, y_M, \hat{y}_M) = \frac{s_{tgt}(y, y_M) - s_{tgt}(y, \hat{y}_M)}{s_{tgt}(y, y_M)} \tag{1}$$

$$d_{src}(x, \hat{x}) = (1 - s_{src}(x, \hat{x}))/1 \tag{2}$$

A successful attack reduces the target side translation similarity more than it reduces the similarity of the perturbed $\hat{x}$ to x. This is reflected in Equation 3, also following Michel et al. (2019) aside from the difference in d_{tgt}. When success, S, is greater than one, the attack achieved that goal. S values below 1 indicate the source side texts were degraded more than the effect on the translation.

$$\begin{aligned} S &= 1 + d_{tgt}(y, y_M, \hat{y}_M) - d_{src}(x, \hat{x}) \\ &= \frac{s_{tgt}(y, y_M) - s_{tgt}(y, \hat{y}_M)}{s_{tgt}(y, y_M)} + s_{src}(x, \hat{x}) \end{aligned} \tag{3}$$

We estimate both source- and target-side similarity using CHRF (Popović, 2015). This metric has been found to be well-correlated to human perception of meaning preservation for varied machine edits (Michel et al., 2019; Merkhofer et al., 2021), and it best matches human perception of machine translation quality at a segment level for the language pairs studied (Mathur et al., 2020).

2.2 Calibrating Meaning Preservation Metrics in Multiple Languages

Most semantic similarity metrics are designed and tested to match human judgments in one language, but they generally aren't calibrated to line up with each other *across languages*. The success criterion in Equation 3 directly compares similarity in source and target languages, but a similarity reduction in the source language needs to be commensurate with the similarity reduction in the target language. Otherwise, perturbations may game the difference between the two scales rather than truly exploiting an MT weakness. A set of examples motivating the differences in CHRF values in different languages can be see in Table 1. This can be replicated for other languages and other metrics, as MT metrics are typically not calibrated across languages, especially not at sentence-level granularity.

Calibration of metrics in different languages relies on common sources of strings with the same distribution of meanings in the two languages. We collect a distribution of $s(x_i, x_j)$ values from random strings following the same sampling pattern in each of the languages. We convert those empirical distributions into complementary cumulative distribution functions (CCDF) to work well with log scale. Figure 1 shows the empirical distribution of CHRF scores accumulated using random strings sourced form WMT20 parallel data (Barrault et al., 2020). The samples were synchronized across the languages so the same underlying distributions were reflected in each curve. Using sampled $i, j \in 1 \ldots N$ from the N sentences available, strings used are $i_{\lfloor x\texttt{len}(i)\rfloor : \lfloor y\texttt{len}(i)\rfloor}$ and $j_{\lfloor z\texttt{len}(j)\rfloor : \lfloor w\texttt{len}(j)\rfloor}$, where $x < y, z < w \in [0, 1)$.

Conversion of the calibrated scores from the CCDFs to a common language and range is done by linear interpolation. Each curve is approximated with 1000 points as shown in Figure 1. To calculate an associated English CHRF$_{EN}$ for an input Chinese CHRF value, x, we find the closest two surrounding x-axis pairs of the *zh* CCDF curve and interpolate between them to get y', the estimated CCDF value for that input x. Then the process is performed again using the *en* curve. We find the two closest y-values on the *en* curve and interpolate using y' to find an x' on the x-axis of the *en* curve. The resulting converted metrics, all calibrated to English, are shown in Figure 2.

σ	CHRF	segment pair
0.21	0.61	Afghanistan boosts security for presidential election A massive security operation is in place across Afghanistan for the country's presidential election on Saturday.
	0.65	Afghanistan verstärkt die Sicherheit für die Präsidentschaftswahlen Für die Präsidentschaftswahlen am Samstag ist in ganz Afghanistan eine umfangreiche Sicherheitsoperati on im Gange.
	0.58	Afghánistán zvyšuje bezpečnostní opatření provázející prezidentské volby V Afghánistánu probíhají masivní bezpečnostní opatření pro zajištění bezpečnosti při sobotních prezide ntských volbách.
	0.13	阿富汗加强安保应对总统选举 阿富汗在全国范围内开展了大规模的安保行动，为星期六的国家总统大选做好准备。
0.29	0.89	Men undergoing surgery for prostate cancer fare as well without radiotherapy Men undergoing surgery for prostate cancer fare just as well without radiotherapy, a major study has found.
	0.48	Männern geht es nach einer Operation wegen Prostatakrebs mit und ohne Strahlentherapie gleich gut Laut einer Studie gibt es keinen Unterschied, ob sich Männer, die wegen Prostatakrebs operiert wurden, einer Strahlentherapie unterziehen oder nicht.
	0.96	Muži, kteří trpí rakovinou prostaty a jdou na operaci, nemusejí podstoupit radioterapii Muži, kteří trpí rakovinou prostaty a jdou na operaci, nemusejí podstoupit radioterapii, zjistila studie.
	0.24	前列腺癌手术后跳过放疗，效果良好 一项重大研究发现，接受前列腺癌手术的男性在不接受放射治疗的情况下状态良好。

Table 1: Examples of high CHRF variance from the WMT20 dataset. σ is the standard deviation of the set of four CHRF scores. Each pair in a set would ideally exhibit the same meaning preservation score.

2.3 Crafting Attacks without Reference Translations

Black-box adversarial examples are crafted by probing the victim system for translations, with the goal of finding a perturbed input that minimizes similarity between the system translation and the reference. In the literature, this is typically the same ground truth reference used for evaluation, but in a practical attack, acquiring a human translation for each segment would be prohibitively slow and expensive. We craft perturbations using the system translation of the original source segment in place of a targeted reference translation to simulate a more realistic, low-information scenario where an adversary doesn't have access to a ground truth. In this case, the probes reveal how system behavior changes but not how translation quality with respect to a human reference translation is affected. This allows us to examine whether simply probing the system can effectively predict perturbations to reduce system performance.

2.4 Transfer: Using one MT System as Proxy Target for Another

We examine *transfer attacks* by measuring the effect of each set of black-box perturbations on the other MT systems. Without direct white-box access to the model gradient or black-box access to repeated probes, transfer attacks rely on the nature of language or implicit similarity between systems. When perturbations succeed against another model, the first system can serve as a proxy to craft attacks on the victim system. Intuitions suggest that transfer attacks are less

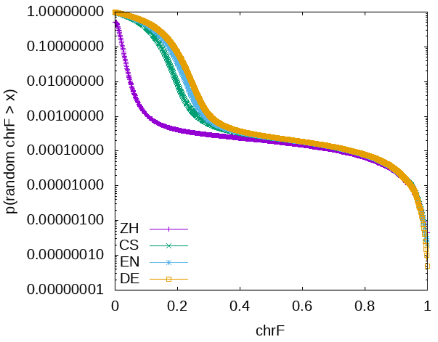

Figure 1: Complementary cumulative distribution functions for CHRF in different languages.

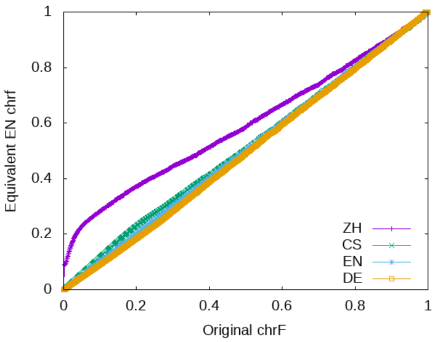

Figure 2: English-equivalent CHRF calibrations.

likely to be effective than black-box attacks, but we want to measure that effect. We also investigate ensemble perturbers, which select edits based on expected performance against multiple MT systems, as an instance of transfer.

3 Experiments

We present a novel attack mechanism that uses a finite state transducer-based paraphraser to generate paraphrases and then selects the best candidate as the attack. We test two targeting conditions, **reference**, where the attack is crafted using ground truth translations from the dataset, and **MT**, where only the system translation of the original source is used. Each set of perturbations is evaluated against the MT system used to craft it (**black box**) and each of the other MT systems (**transfer**). Much of the prior work uses reference translations for attack crafting and presents primarily black-box evaluations, but crafting perturbations using only MT outputs and testing transfer to inaccessible systems is a more realistic, low-information scenario. We compare our adversarial FST to the black-box SEQ2SICK approach from (Cheng et al., 2020) under the same conditions. The evaluation metrics are calibrated CHRF, converted into English-equivalent CHRF$_{EN}$, and Success using calibrated CHRF.

3.1 Data

Our experiments use the WMT 2020 test sets for EN-DE, EN-CS and EN-ZH (Barrault et al., 2020). The source for each target language test set consists of the same 1418 English-language segments from the news domain.

3.2 MT Systems

Our experiments probe eight trained machine translation systems acquired from the Transformers model zoo (HuggingFace, 2020). **mBART** English-to-Many is a transformer with multilingual pretraining that is fine tuned to translate from English into many other languages including DE, CS and ZH (Tang et al., 2020). We use separate bilingual EN-DE, EN-CS and EN-ZH models from **OPUS-MT** (Tiedemann and Thottingal, 2020). We use EN-DE bilingual models from Kasai et al. (2020) (**Allen**) and Ng et al. (2019) (**Facebook**). For every system, we use a beam size of five.

3.3 Attack: Finite State Transducer-based Paraphraser With Rescoring

We produce adversarial examples by first generating a large portfolio of paraphrases $\hat{X} = \{\hat{x}_{1\ldots n}\}$ for the input x, then selecting the best candidate under a configurable mix of source similarity and attack effectiveness. For these experiments, we weight these two factors equally.

To preserve the semantics of an input, our method begins with high-quality paraphrases. We compile 2.3 million *equivalence* paraphrases from the Penn Paraphrase Database (PPDB) (Ganitkevitch et al., 2013) into a finite state transducer (FST) rewriting input strings. We use a log transform of the PPDB2 score, the estimate of human acceptance included with each PPDB entry, as the weight for the transduction and follow the methods of Stahlberg et al. (2019) to generate a lattice of alternatives for input strings. We minimize the lattice FST resulting from the composition of the input string with the transducer, remove epsilons and determinize, then use the shortest path search. We keep the n-best list of alternatives to use as our candidate edit pool, with $n = 1000$. It takes an average of 0.085 seconds on one CPU to obtain 1000 alternatives for the input sentences studied in this paper using the `pynini` toolkit (Gorman, 2016) built on top of OpenFST (Allauzen et al., 2007).

We select one perturbation $\hat{x}$ per segment per system that balances attack effectiveness and meaning preservation. We estimate both in terms of similarities as measured by CHRF. Meaning preservation is measured by comparing the original source and the candidate paraphrase, $s_{src}(x, \hat{x})$. For every translation system M, we obtain translations y_M for x, the original source, and $\hat{y}_M$ for each $\hat{x}$. Attack effectiveness is estimated by measuring how much the system output differs from the output on the original text, that is $s_{tgt}(y_M, \hat{y}_M)$, for the MT condition, or by measuring degradation in translation quality $s_{tgt}(y, \hat{y}_M)$ for the reference condition. For MT system M, we select the candidate $\hat{x}$ that maximizes $f(\hat{x}, M) = s_{src}(x, \hat{x}) - s_{tgt}(y_M, \hat{y}_M)$, equally weighting source attack effectiveness and meaning preservation.

Meaning preservation $s_{src}(x, \hat{x})$ and attack scores $s_{tgt}(y_M, \hat{y}_M)$ are scaled prior to selection with a simple Gaussian transform to get them into a comparable range. Without the score transform, the source language similarity scores would tend to be very high compared to the MT similarity scores. This rescaling makes the aggregate optimization more well-balanced.

3.4 Ensemble attacks

Ensemble attacks were crafted by evaluating attack candidates using multiple MT systems and averaging the resulting target similarity values, $s_{tgt}(y_M, \hat{y}_M)$, when performing the attack selection. **Mean** refers to attacks using all eight systems, $f_{mean}(\hat{x}) = \sum_i f(\hat{x}, M_i)$. Leave-one-out (denoted **loo**) refers to averaging all but the victim system's similarity estimate, $f_{loo}(\hat{x}, M_j) = \sum_{i \neq j} f(\hat{x}, M_i)$ where the victim system is M_j. The leave-one-out condition simulates attacking an otherwise unknown, inaccessible MT system. Both ensemble techniques realized gains in transfer success count as more systems were included in the ensemble.

3.5 Baseline Attack: SEQ2SICK

We use the black-box implementation of SEQ2SICK (Cheng et al., 2020) in the TextAttack python library (Morris et al., 2020b) as a baseline attack on machine translation. This targeted attack generates candidate edits by swapping words for other words that are close in word embedding space. It obtains translations for each candidate from the model and greedily applies one-word changes that minimize the number of words that are present in both the reference and the translation. For the MT condition, we treat the translation of the original source as the reference.

Using a GPU, one attack takes an average of 32 seconds and 285 probes when targeting the reference or 35 seconds and 313 probes when targeting the system translation of the original source.

			Uncal.		Calibrated	
		N	Ref	MT	Ref	MT
FST+Rerank	black box	8	8	8	8	8
	transfer	56	31	9	16	2
	mean ensemble	8	8	8	8	7
	loo ensemble	8	8	6	7	4
S2S	black box	8	2	2	1	1
	transfer	56	16	20	3	8

Table 2: Effects of calibration and reference access at crafting time on black box and transfer success counts. Uncalibrated and reference-crafted configurations overestimate success.

		Ref S			MT S		
		>1	=1	<1	>1	=1	<1
FST+Rerank	black box	65	35	0	66	1	33
	transfer	26	35	39	35	1	64
	mean ensemble	21	68	11	51	1	48
	loo ensemble	16	66	18	42	1	67
S2S	black box	30	0	70	34	0	66
	transfer	28	0	72	31	0	69

Table 3: Percent of sentences for which S was > 1 (successful), $= 1$ (no viable attack found), < 1 (unsuccessful). Crafted with access to reference (Ref) and without (MT), calibrated conditions only.

4 Results

Success Every set of perturbations degrades the translation quality of every model. All sets of black-box perturbations using the FST-based perturber meet the criterion of success under both targeting conditions. However, many transferred FST-based attacks and many SEQ2SICK attacks under both conditions do not achieve success. Table 2 counts the number of successes over both perturbers under different conditions. Table 4 presents more details for attacks using the FST-based perturber, which is more often successful. Each system's performance on the unperturbed WMT20 dataset, as measured for this study, is reported as original CHRF.

Effects of Calibration Table 2 shows the effect calibration has on success rates. Tuning and measuring performance with calibrated metrics reveals that uncalibrated metrics overestimate success. The systems in the uncalibrated conditions exploited mismatches in the CHRF scales for different languages rather than vulnerabilities of the MT systems.

Referenceless attacks Attacks crafted against the reference achieve a higher margin of success under black-box scenarios and are much more likely to transfer than attacks crafted against the original system output. This suggests that the changes made under the MT condition are more tailored to the errors in the system with which they were crafted, perhaps by further changing parts of the system translation that already do not match the ground truth. Since transfer tends to reduce adversarial effect, the effect of these weaker attacks less frequently outweighs the degradation of the source.

Attack transfer These results don't suggest trends in transfer that correspond to system/language similarity or relative performance. While Allen and Opus-DE were relatively vulnerable to reference-targeted attacks from other systems, this vulnerability doesn't extend to MT-targeted attacks. Adversaries crafted using the multilingual model, mBART, do not transfer better between its different target languages, even though they share model weights.

Ensemble attacks often transfer: The mean ensembles are successful against nearly all individual models and the leave-one-out ensemble attacks successfully transfer in eleven of the sixteen settings. Continuing to add more MT systems to a leave-one-out targeting system would likely increase its effectiveness. Favoring attacks that succeed against more targeting systems leads to better transfer to previously unseen systems.

Example 1: Changing translated day of week
(mBART MT, $s_{src} = 0.94, d_{tgt} = 0.32$)

original	Sacramento police also announced Thursday their internal investigation did not find any policy or training violations.
attack	Sacramento police also announced Today (thursday) their internal investigation did not find any policy or training violations.
reference	Die Polizei von Sacramento gab am Donnerstag ebenfalls bekannt, dass ihre innere Ermittlung keine Verletzung der Regeln oder des Trainings erkennen ließ.
original output	Sacramento Polizei gab auch am Donnerstag [Thursday] bekannt, dass ihre interne Untersuchung keine Verstöße gegen die Richtlinien oder Ausbildung gefunden hat.
attack output	Sacramento Polizei auch angekündigt Heute (Samstag) [Saturday] ihre interne Untersuchung fand keine Politik oder Ausbildung Verletzungen.

Example 2: Omitting the object in perturbed translation
(Facebook MT, $s_{src} = 0.91, d_{tgt} = 0.32$)

original	Prince Harry detonated a recently detected mine in Angola.
attack	Prince Harry detonated most recently detected mine in Angola.
reference	Prinz Harry detonierte eine kürzlich entdeckte Mine in Angola.
original output	Prinz Harry hat eine kürzlich entdeckte Mine in Angola gesprengt.
attack output	Prinz Harry detonierte zuletzt in Angola. [Prince Harry last detonated in Angola.]

Example 3: Unrelated translation
(Allen MT, $s_{src} = 0.99, d_{tgt} = 0.89$)

original	Many readers, including some who work in national security and intelligence, have criticized The Times's decision to publish the details, saying it potentially put the person's life in danger and may have a chilling effect on would-be whistle-blowers.
attack	Many readers, including some who work in national security and intelligence, have criticized The Times's decision to publish the details, 's saying it potentially put the person's life in danger and may have a chilling effect on would-be whistle-blowers.
reference	Vieler Leser, darunter auch einige, die für die nationalen Sicherheits- und Nachrichtendienste arbeiten, haben die Entscheidung von The Times, Details zu veröffentlichen, kritisiert und geäußert, dass dadurch wahrscheinlich das Leben der Person in Gefahr gebracht wurde und es einen abschreckenden Effekt auf potenzielle Whistleblower haben könnte.
original output	Viele Leser, darunter einige, die in national Sicherheit und Intelligenz arbeiten, haben die Entscheidung der Times kritisiert, die Details zu veröffentlichen, sagte, dass sie potenziell das Leben der Person in danger und könnte eine abschreckende Wirkung auf würde -@ be whistle -@ be whistle -@ blowers.
attack output	Die Times ist eine US-amerikanische Schauspielerin. [The Times is an American actress.]

Figure 3: Examples with perturbations in orange and back translations in blue.

					English-Czech		English-German				English-Chinese	
					mbart	opus	allen	fb	mbart	opus	mbart	opus
original s_{tgt}, CHRF					53.9	54.8	46.7	63.9	58.2	60.0	27.9	26.1
calibrated s_{tgt}, CHRF$_{en}$					54.5	55.3	45.7	63.3	57.5	59.3	42.6	41.1
selector			$\mathcal{L}_{src}$	s_{src}↑	Success, S ↑							
Crafted with Reference	CS	mbart	1.23	97.1	**1.12**	**1.02**	**1.00**	0.99	1.00	1.00	0.98	0.99
		opus	1.33	96.8	**1.01**	**1.14**	**1.01**	0.99	1.00	**1.00**	0.98	0.99
	DE	allen	1.53	96.1	0.99	0.99	**1.27**	0.99	1.00	**1.01**	0.97	0.98
		fb	1.44	96.3	1.00	1.00	**1.02**	**1.13**	**1.01**	**1.02**	0.98	0.99
		mbart	1.29	96.9	1.00	1.00	**1.02**	1.00	**1.12**	**1.02**	0.98	0.99
		opus	1.36	96.6	1.00	**1.00**	**1.02**	**1.00**	**1.02**	**1.13**	0.98	0.99
	ZH	mbart	0.84	98.3	1.00	1.00	1.00	1.00	1.00	1.00	**1.10**	**1.00**
		opus	1.01	97.7	0.99	1.00	1.00	1.00	1.00	1.00	1.00	**1.14**
		mean	0.51	99.2	**1.02**	**1.02**	**1.07**	**1.02**	**1.02**	**1.02**	**1.01**	**1.02**
		loo	0.47	$\sim$ 99.1	**1.01**	**1.01**	**1.01**	**1.01**	**1.01**	**1.01**	1.00	**1.00**
Crafted with MT	CS	mbart	2.61	92.8	**1.06**	**1.00**	0.99	0.98	0.99	0.99	0.96	0.97
		opus	2.61	92.7	0.99	**1.08**	0.99	0.97	0.98	0.99	0.95	0.97
	DE	allen	2.66	92.2	0.97	0.98	**1.20**	0.97	0.98	0.99	0.95	0.96
		fb	2.61	92.6	0.98	0.99	1.00	**1.07**	0.99	1.00	0.96	0.97
		mbart	2.52	93.0	0.99	0.99	1.00	0.98	**1.07**	**1.00**	0.96	0.97
		opus	2.56	93.0	0.98	0.99	1.00	0.97	0.99	**1.08**	0.96	0.97
	ZH	mbart	2.49	93.1	0.98	0.98	0.99	0.97	0.98	0.99	**1.04**	0.98
		opus	2.52	93.2	0.98	0.99	0.98	0.97	0.98	0.99	0.97	**1.08**
		mean	2.67	93.7	**1.01**	**1.03**	**1.08**	**1.01**	**1.02**	**1.03**	0.99	**1.02**
		loo	2.67	$\sim$ 93.6	1.00	**1.01**	**1.02**	0.99	**1.00**	**1.01**	0.97	0.99

Table 4: Full FST+Rerank targeted attack results using calibrated metrics. Successful attacks in bold. **mean** and **loo** represent targeting with the mean and leave-one-out ensembles. s_{src} measures meaning preservation on the source side using CHRF. s_{tgt} is the MT score of the output sentence, measured with CHRF calibrated to English. $\mathcal{L}_{src}$ is the mean edit distance between attacks and originals. Note reference-informed attacks exhibit more conservative edits.

Examples The examples in Figure 3 illustrate some of the range of translation errors given perturbed inputs. They are drawn from black-box FST-based perturbations against the four EN-DE translation systems.

5 Related Work

The performance of machine translation systems is vulnerable to adversarial examples of several types. Naturalistic and untargeted changes degrade system performance, while remaining largely intelligible to humans (Belinkov and Bisk, 2018). Using word- or character-level permutations, untargeted attacks simply degrade translation quality, while targeted attacks introduce particular errors such as removing or inserting selected words. White-box attacks perturb an input with access to the model's gradients (Ebrahimi et al., 2018; Cheng et al., 2019; Wallace et al., 2019; Cheng et al., 2020) while a black-box paradigm only probes the model's output, typically for salience of portions of the input and scoring of substitutions proposed via heuristics (Zhao et al., 2018; Zhang et al., 2021). Other work crafts attacks based on generally exploitable features of language that are discoverable in training data, such as polysemous words, without probing expected attack success (Emelin et al., 2020).

Adversarial examples are crafted with respect to particular models and challenge datasets and they achieve limited success when applied (transferred) to others. A range of text classification adversaries have been shown to reduce the accuracy of models that have different architectures or were trained on different datasets (Song et al., 2021; Ren et al., 2019; Song et al., 2020; Emmery et al., 2021). While transfer effectiveness varies by attack method, it does not reach the level of the matched condition. Several authors show that their adversarial examples, created using white-box attacks on known systems, transfer to some extent to publicly available APIs hosted by Google, Baidu and Bing (Zhao et al., 2018; Zhang et al., 2021; Gil et al., 2019). Emelin et al. (2020) find that their attacks based on dataset co-occurrence reduce the accuracy of several models, but there's little overlap in which examples succeed, with slightly more similarity in sets of examples that are successful on models with the same architecture. White-box, gradient-based attacks can be crafted on models "stolen" via knowledge distillation, despite mismatches in data domain and model architecture (Wallace et al., 2020).

Adversarial perturbations typically must conform to perceptual features of an original text. Most NLP attack methods apply one-off perceptual constraints or preferences (e.g. lower number of swaps or similarity among vector representations) but the tradeoff between attack effectiveness and human perception is often unaccounted for, making it difficult to discern when an adversarial effect is the result of perturbations that are easily detected by a human observer (Morris et al., 2020a). Michel et al. (2019) propose a metric for success that balances adversarial effect with the level of meaning preservation of the original.

Paraphrases have recently been used for improving evaluation of MT (Bawden et al., 2020; Thompson and Post, 2020a), for improving MT training (Khayrallah et al., 2020) and multitask MT models have been run in a clever way to generate paraphrases (Thompson and Post, 2020b). The adversarial inputs of Iyyer et al. (2018) are generated using a neural end-to-end paraphrase system.

6 Conclusion

In this paper, we considered the practicality of adversarial examples for NLP by crafting MT attacks without access to the victim system or ground truth and by measuring those attacks in a way that accounts for both attack effectiveness and source meaning preservation. We find that many attacks that reduce translation quality still fall short of a strict threshold of *success*. We investigated the ability to transfer attacks across systems and across MT target languages. Attacks that do not have access to ground truth rarely transfer between systems. When they are crafted using ground truth, they transfer more often but we did not observe patterns, like language or system similarity, that allow us to predict when transfer will occur.

Our FST perturbation process is able to select edits under configurable constraints that preserve source-side meaning while causing large changes in system output. This is due in part to a high-quality paraphrase generation process relying on millions of paraphrases with scores calibrated to human quality judgments. This selection process is sufficient to degrade translation quality with respect to ground truth. The construction of candidates and attack selection processes do not require a GPU. Ensembles performed the highest rate of successful attacks.

One direction for future work could investigate methods for improving system robustness to attacks of this type. The leave-one-out ensemble was the most reliable attack method we found with at least 50% success rate in all conditions, including transferring attacks to systems it had no previous access to. Building on that success, cultivating it to a robust attack mechanism spanning languages and systems could be another valuable contribution in the future.

Ethical Considerations

There is a risk that adversarial techniques will be used by malicious actors to attack real world NLP systems. We believe that sharing this knowledge allows people who deploy models to account for risk and create safer systems; in particular, we examine how effectiveness measures and techniques reported in recent literature might look under more practical, low-information scenarios outside of academic test harnesses.

Our work is part of a thread in AI assurance that uncovers vulnerabilities and feeds research into mitigation methods, such as model robustness and detection of deceptive inputs.

Acknowledgements

Portions of this technical data were produced for the U. S. Government under Contract No. FA8702-19-C-0001 and W56KGU-18-D-0004, and is subject to the Rights in Technical Data-Noncommercial Items Clause DFARS 252.227-7013 (FEB 2014).

References

Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., and Mohri, M. (2007). Openfst: A general and efficient weighted finite-state transducer library. In *International Conference on Implementation and Application of Automata*, pages 11–23. Springer.

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C.-k., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M., and Zampieri, M. (2020). Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Bawden, R., Zhang, B., Tättar, A., and Post, M. (2020). ParBLEU: Augmenting metrics with automatic paraphrases for the WMT'20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 887–894, Online. Association for Computational Linguistics.

Belinkov, Y. and Bisk, Y. (2018). Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Cheng, M., Yi, J., Chen, P.-Y., Zhang, H., and Hsieh, C.-J. (2020). Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3601–3608.

Cheng, Y., Jiang, L., and Macherey, W. (2019). Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4324–4333, Florence, Italy. Association for Computational Linguistics.

Ebrahimi, J., Lowd, D., and Dou, D. (2018). On adversarial examples for character-level neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 653–663, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Emelin, D., Titov, I., and Sennrich, R. (2020). Detecting word sense disambiguation biases in machine translation for model-agnostic adversarial attacks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7635–7653, Online. Association for Computational Linguistics.

Emmery, C., Kádár, Á., and Chrupała, G. (2021). Adversarial stylometry in the wild: Transferable lexical substitution attacks on author profiling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2388–2402, Online. Association for Computational Linguistics.

Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013). PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia. Association for Computational Linguistics.

Gil, Y., Chai, Y., Gorodissky, O., and Berant, J. (2019). White-to-black: Efficient distillation of black-box adversarial attacks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1373–1379, Minneapolis, Minnesota. Association for Computational Linguistics.

Gorman, K. (2016). Pynini: A python library for weighted finite-state grammar compilation. In *Proceedings of the SIGFSM Workshop on Statistical NLP and Weighted Automata*, pages 75–80.

HuggingFace (2020). Hugging face model zoo. Accessed: 2021-03.

Iyyer, M., Wieting, J., Gimpel, K., and Zettlemoyer, L. (2018). Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of NAACL*.

Kasai, J., Pappas, N., Peng, H., Cross, J., and Smith, N. (2020). Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation. In *International Conference on Learning Representations*.

Khayrallah, H., Thompson, B., Post, M., and Koehn, P. (2020). Simulated multiple reference training improves low-resource machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 82–89, Online. Association for Computational Linguistics.

Mathur, N., Wei, J., Freitag, M., Ma, Q., and Bojar, O. (2020). Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Merkhofer, E., Mendoza, M.-A., Marvin, R., and Henderson, J. (2021). Perceptual models of machine-edited text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3909–3920, Online. Association for Computational Linguistics.

Michel, P., Li, X., Neubig, G., and Pino, J. (2019). On evaluation of adversarial perturbations for sequence-to-sequence models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3103–3114, Minneapolis, Minnesota. Association for Computational Linguistics.

Morris, J., Lifland, E., Lanchantin, J., Ji, Y., and Qi, Y. (2020a). Reevaluating adversarial examples in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3829–3839, Online. Association for Computational Linguistics.

Morris, J., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., and Qi, Y. (2020b). TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., and Edunov, S. (2019). Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ren, S., Deng, Y., He, K., and Che, W. (2019). Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.

Song, C., Rush, A., and Shmatikov, V. (2020). Adversarial semantic collisions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4198–4210, Online. Association for Computational Linguistics.

Song, L., Yu, X., Peng, H.-T., and Narasimhan, K. (2021). Universal adversarial attacks with natural triggers for text classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3724–3733, Online. Association for Computational Linguistics.

Stahlberg, F., Bryant, C., and Byrne, B. (2019). Neural grammatical error correction with finite state transducers. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4033–4039, Minneapolis, Minnesota. Association for Computational Linguistics.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning.

Thompson, B. and Post, M. (2020a). Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Thompson, B. and Post, M. (2020b). Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570, Online. Association for Computational Linguistics.

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Wallace, E., Feng, S., Kandpal, N., Gardner, M., and Singh, S. (2019). Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Wallace, E., Stern, M., and Song, D. (2020). Imitation attacks and defenses for black-box machine translation systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5531–5546, Online. Association for Computational Linguistics.

Zhang, X., Zhang, J., Chen, Z., and He, K. (2021). Crafting adversarial examples for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1967–1977, Online. Association for Computational Linguistics.

Zhao, Z., Dua, D., and Singh, S. (2018). Generating natural adversarial examples. In *International Conference on Learning Representations*.