Sentiment Perception Adversarial Attacks on Neural Machine Translation Systems

Vyas Raina [1] **Mark Gales** [1]

Abstract

With the advent of deep learning methods, Neural Machine Translation (NMT) systems have become increasingly powerful. However, deep learning based systems are susceptible to adversarial attacks, where imperceptible changes to the input can cause undesirable changes at the output of the system. To date there has been little work investigating adversarial attacks on sequence-to-sequence systems, such as NMT models. Previous work in NMT has examined attacks with the aim of introducing target phrases in the output sequence. In this work, adversarial attacks for NMT systems are explored from an output perception perspective. Thus the aim of an attack is to change the perception of the output sequence, without altering the perception of the input sequence. For example, an adversary may distort the sentiment of translated reviews to have an exaggerated positive sentiment. In practice it is challenging to run extensive human perception experiments, so a proxy deep-learning classifier applied to the NMT output is used to measure perception changes. Experiments demonstrate that the sentiment perception of NMT systems' output sequences can be changed significantly with small imperceptible changes to input sequences. Link to code: `https://github.com/rainavyas/SentAttackNMT`

1. Introduction

Deep learning based Neural Machine Translation (NMT) systems are used ubiquitously for automatic translation of texts. However, deep learning based systems are susceptible to adversarial attacks (Szegedy et al., 2013), where small imperceptible changes at the input of the system can result in significant, undesired, changes at the output. In the natural language domain, many papers (Sun et al., 2018; Cheng et al., 2018; Blohm et al., 2018; Neekhara et al., 2018; Jia & Liang, 2017; Iyyer et al., 2018; Zhao et al., 2017; Raina et al., 2020) have identified methods to generate adversarial examples. To date most works have focused on text classification: the aim is to alter the textual input such that the system mis-classifies.

NMT systems, however, perform a sequence-to-sequence (S2S) task, where an input, source text sequence is mapped to an output target text sequence, which for NMT is the translation of the source. The definition of an attack needs to be modified for these S2S tasks. Cheng et al. (2018) introduces the concept of non-overlapping attacks (output sequence should be completely changed) and target keyword attacks (insert target words in the output sequence). Ebrahimi et al. (2018); Zou et al. (2019); Zhang et al. (2021) describe methods to perform target keyword attacks specifically for NMT systems. Through the use of a new evaluation framework for S2S adversarial perturbations, Michel et al. (2019) reveal that many existing methods do not preserve semantic meaning - hence they modify these attacks with added constraints. Although this gives a realistic setting for many adversarial attacks, it does not capture attacks that seek to change the *perception* of the output sequence (e.g. Raina et al. (2022)). An adversary may, for example, want to change the input text (in an imperceptible manner) such that the output text reads negatively to a human reader, without the content of the translation actually changing, e.g. an attack may cause an output sequence *I won the competition* to become *I hardly won the competition*. This form of attack is of concern in many automated translation settings, e.g. product reviews when translated are exaggeratedly positive to attract customers or conversely a benign social media post in one language (not flagged by any detectors) translates to generate negative sentiment hate speech.

To the best of our knowledge, the exploration of adversarial attacks specifically targeting the output perception of sequence-to-sequence (S2S) systems has not been previously undertaken. As such, the primary contribution of this research lies in the expansion and generalization of the defi-

[1]Machine Intelligence Lab, University of Cambridge, United Kingdom. Correspondence to: Vyas Raina <vr313@cam.ac.uk>, Mark Gales <mjfg@eng.cam.ac.uk>.

2nd AdvML Frontiers workshop at 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.
Copyright 2023 by the author(s).

nition of adversarial attacks for S2S systems, encompassing attacks that aim to manipulate the perception of the system's output. In order to showcase the viability of this novel form of attack, a series of experiments were conducted to effectively alter the sentiment perception in the output generated by various NMT systems. Given the costly and impractical nature of conducting extensive human evaluations for perception analysis, this study employs state-of-the-art sentiment classifiers with high performance levels as reliable proxies for measuring perception changes. Representative human evaluation experiments are conducted to verify the validity and appropriateness of the selected proxy classifiers. By leveraging these proxy classifiers, this work provides valuable insights into the alteration of output perception in S2S systems, shedding light on the need for further investigation and the development of effective defenses against such attacks.

2. Perception-Based Adversarial Attacks

Sequence-to-sequence models, with parameters θ, map a T-length input sequence, $x_{1:T}$, to a $\hat{L}$-length output word sequence, $\hat{y}_{1:\hat{L}}$,

$$\hat{y}_{1:\hat{L}} = \mathcal{F}_\theta(x_{1:T}) = \arg\max_{y_{1:L}} \{p(y_{1:L}|x_{1:T};\theta)\} \quad (1)$$

A perception-based adversarial attack aims to generate an adversarial example, $\tilde{x}_{1:\tilde{T}}$, that is mapped to the output sequence $\mathcal{F}_\theta(\tilde{x}_{1:\tilde{T}})$ where the "perception" of this output sequence has changed,

$$\phi(\mathcal{F}_\theta(\tilde{x}_{1:\tilde{T}})) \neq \phi(\mathcal{F}_\theta(x_{1:T})). \quad (2)$$

Here $\phi()$ is a proxy function that mimics human perception of the output. For example the perception could be how positive a sequence is, thus $\phi()$ would be a sentiment classifier. It is necessary for the adversarial attack to satisfy an imperceptibility constraint, $\mathcal{G}()$, which should again mimic human perception. Thus $\mathcal{G}(x_{1:T}, \tilde{x}_{1:\tilde{T}}) \leq \epsilon$, where ϵ is the threshold of imperceptibility. It is difficult to define an appropriate function $\mathcal{G}()$ for word sequences. Perturbations can be measured at a character, word or sentence level. Alternatively, the perturbation could be measured in the vector embedding space, using for example l_p-norm based (Goodfellow et al., 2015) metrics or cosine similarity (Carrara et al., 2019). However, constraints in the embedding space do not guarantee human imperceptibility in the original word sequence space. To ensure the adversarial input sequence, $\tilde{x}_{1:\tilde{T}}$ is visually, semantically and perceptively similar to the original input sequence $x_{1:T}$, this work defines the imperceptibility constraint using four measures:

1. *Perception similarity* - A perception-based adversarial attack should not significantly alter the perception of the input sequence. This can be measured using the same proxy function as used for the output sequence in Equation 2,

$$|\phi(\tilde{x}_{1:\tilde{T}}) - \phi(x_{1:T})| \leq \epsilon_1. \quad (3)$$

2. *Visual similarity* - A normalised variant of a Levenshtein, *edit-based* measurement (Li et al., 2018) is used to limit visual changes,

$$\frac{1}{T}\mathcal{L}(x_{1:T}, \tilde{x}_{1:\tilde{T}}) \leq \epsilon_2, \quad (4)$$

where $\mathcal{L}()$ counts the number of changes between the original sequence, $x_{1:T}$ and the adversarial sequence $\tilde{x}_{1:\tilde{T}}$, where a change is a swap/addition/deletion. This is a standard approach to ensure that adversarial examples do not deviate visually from real examples.

3. *Perplexity* - Adversarial examples should not be easily detectable by automatic detectors. Small changes can result in incomprehensible phrases that can be easily detected using perplexity as calculated by a standard language model (LM) (Raina & Gales, 2022). Hence, the changes should limit perplexity of the adversarial sequence,

$$p_{\text{LM}}(\tilde{x}_{1:\tilde{T}}) \leq \epsilon_3. \quad (5)$$

4. *Semantic Similarity* - The meaning/content of a sentence should not change significantly. Character level attacks are not considered in this work, as they can be easily detected using spelling and grammatical checks (Sakaguchi et al., 2017). Attacks that substitute $N = \epsilon_2 T$ words (recall ϵ_2 is the maximum fraction of edits permitted by the imperceptibility constraint in Equation 4) are considered. As an example, for an input sequence of T words, a N-word substitution adversarial attack, $\tilde{x}_{1:N}$, applied at word positions $n_1, n_2, \ldots, n_N$ gives the adversarial sequence, $\tilde{x}_{1:\tilde{T}}$

$$\begin{aligned}\tilde{x}_{1:\tilde{T}} = x_1, \ldots, x_{n_1-1}, \tilde{x}_1, x_{n_1+1}, \ldots, \\ x_{n_N-1}, \tilde{x}_N, x_{n_N+1}, \ldots, x_T.\end{aligned} \quad (6)$$

It is necessary to select which words to replace, and what to replace them with. As suggested by Ren et al. (2019), a simple approach is to use saliency to rank the word positions in $x_{1:T}$. The N most salient words are then substituted. As the aim is to change the sentiment perception of the output sequence (Equation 2), a modified version of saliency is considered, *sentiment saliency*. For each word, x_t, in sequence $x_{1:T}$ this is defined as

$$\begin{aligned}\mathcal{S}(x_t|x_{1:T}) = |\phi(\mathcal{F}_\theta(x_{1:t-1}, x_{t+1:T})) \\ - \phi(\mathcal{F}_\theta(x_{1:T}))|.\end{aligned} \quad (7)$$

To ensure small semantic changes, **only word synonyms** are considered for the substitutions.

The attack method described in this section uses two proxy functions to allow for automatic adversarial example generation at scale. Four imperceptibility measures are introduced and then a proxy sentiment classifier, $\phi()$, is used to measure output sentiment. To validate the use of these proxy measures/functions, human evaluation experiments are conducted (results in Section 3.3).

3. Experiments

3.1. Experimental Setup

Experiments are performed using the NMT data from the WMT19 news translation task (Foundation). Results are presented for the Russian (ru) to English (en), German (de) to English and reverse translation tasks, where there are 2000 test examples. The best models, submitted by FAIR (Ng et al., 2019), are used as the baseline[1]. Table 1 gives the performance of these models on the WMT19 test set (respectively for each language pair), calculated using the SacreBleu tool (Post, 2018).

Task	BLEU	CHRF	TER
de-en	41.20	65.11	47.66
ru-en	38.81	63.37	49.73
en-de	42.77	67.55	46.85
en-ru	38.81	63.37	49.73

Table 1. Model performances on WMT19 test sets

3.2. Attack Results

Each translation model is attacked using the saliency-based synonym substitution attack described in Equation 6, where the aim is to increase either the *positivity* or *negativity* sentiment of the output text sequences. The imperceptibility constraint on perplexity in Equation 5 is enforced by ensuring the sentence perplexity, as measured by popular GPT2-based language models [2] (Radford et al., 2019), is less than 1.5 times the average (across dataset) sentence perplexity. This constraint cannot be too strict as in many applications (e.g. tweets/product reviews) authentic sentences contain many grammatical errors. Additionally, to satisfy Equation 3, the sentiment perception (positive/negative classification) of the input sequence is constrained to not change. As a proxy for human sentiment perception (human evaluation experiments in Section 3.3.2), sentiment classification of input and output text sequences of the translation models is measured using standard pre-trained Transformer-based (Vaswani et al., 2017) sentiment classifiers: the sentiment of the English sequences is measured using a pre-trained (on 58M tweets) Roberta based sentiment classifier[3]; the sentiment of Russian sequences is measured using RuBERT, a pre-trained Russian BERT system[4]; and the sentiment of German sequences is measured using a Bert Based German sentiment classifier, pretrained on texts from Twitter, Facebook and app reviews (1.83M samples) [5]. The candidate list of synonyms for substitution are found using popular NLP tools for each language: the `wordnet` lexical database (Fellbaum, 1998) is used for English; `wiki-ru-wordnet` tool (wiki-ru wordnet) for Russian; and the `OdeNet` tool (odenet) for German. Table 2 gives an example of an attack on the *de-en* NMT system.

Figure 1 shows the impact (*NMT attack* curves) of the adversarial attacks of increasing strength (fraction of words substituted, as defined by imperceptibility constraint Equation 4), measured by the percentage of test samples classified as positive/negative[6]. Results here are presented for the ru-en and de-en systems[7]. It is interesting to observe that language pairing has a strong influence on the impact of the sentiment attacks. For example, in both the negative and positive attack scenarios, the de-en NMT system has more than a 30% increase in fraction of samples with positive/negative sentiment, whilst for the ru-en NMT system's increase is limited to less than 20%. Nevertheless, these results demonstrate that both NMT systems are susceptible to attacks where significant changes in output sequences' sentiment perception can be achieved with imperceptible changes at the input.

Figure 1 gives one further curve for each NMT system: *direct attack*. Here, the same synonym substitution attack approach of Equation 6 is used to directly attack the predicted output English sequence [8] to increase the positive/negative sentiment score predicted by the English sentiment classifier. The substitutions are again limited to word synonyms and the perplexity constraint of Equation 5 is enforced. Note that this direct attack is presented only as a point of comparison, as in the attack of a NMT system an adversary realistically

[1] All NMT trained models available at: https://huggingface.co/facebook/wmt19-de-en.

[2] Perplexity LM Models: English-https://huggingface.co/distilgpt2; German-https://huggingface.co/dbmdz/german-gpt2; Russian-https://huggingface.co/sberbank-ai/rugpt3large_based_on_gpt2.

[3] English sentiment classifier available at: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

[4] Russian sentiment classifier available at: https://huggingface.co/blanchefort/rubert-base-cased-sentiment-rusentiment

[5] German sentiment classifier available at: https://huggingface.co/oliverguhr/german-sentiment-bert

[6] Predictions are made using a max-class classification rule.

[7] Results for the en-de and en-ru NMT systems are given in Figure 2

[8] Identical trends were found when instead of the predicted sequences the reference English sequences were *directly* attacked.

	Original	Attacked
Source	Neun Minuten vor Schluss buxierte Watford-Verteidiger Craig Cathcart eine Hereingabe von Alex Iwobi unglücklich ins eigene Tor, nur zwei Minuten später sorgte Mesut Özil mit seinem dritten Saisontreffer für die Entscheidung.	Neun Minuten vor Ausgang buxierte Watford-Verteidiger Craig Cathcart eine Hereingabe von Seiten Alex Iwobi deplorabel ins eigene Tor, nur zwei Minuten später sorgte Mesut Özil mit seinem dritten Saisontreffer für die Beschluss.
Prediction	Nine minutes from the end Watford defender Craig Cathcart unluckily booked an own goal from Alex Iwobi, and just two minutes later Mesut Özil secured the win with his third goal of the season.	Nine minutes from time Watford defender Craig Cathcart netted an own goal from Alex Iwobi, and just two minutes later Mesut Özil made sure with his third goal of the season.
Sentiment	**8%** positive	**61%** positive

Table 2. Adversarial attack example on de-en NMT system. Target *positive* sentiment.

	Original	Attacked
Source	Данный договор должен решить не только многолетний спор о названии страны, но и открыть Скопье путь в НАТО и ЕС.	Данный сделка должен решить не всего многолетний спор о названии страны, но и открыть Скопье путь во благо НАТО и ЕС.
Prediction	The treaty should resolve not only the long-standing name dispute, but also open Skopje's path to NATO and the EU.	The deal should not only resolve the long-standing name dispute, but also pave the way for Skopje to benefit NATO and the EU.
Sentiment	21% positive	48% positive

Table 3. Adversarial attack examples on ru-en NMT system. Target positive sentiment.

only has access to the source text. For the *ru-en* system, in both the positive and negative attack settings, as would be expected, the direct attack of the sentiment classifier gives an upper-bound to the indirect NMT attack. However, the indirect *NMT attack*, in the positive and negative attack settings on the *de-en* NMT system, is more powerful for up to 40% words substituted, than the *direct attack* on the English sentiment classifier. This suggests that an attack on the NMT system can generate an output sequence (in English) that is in fact more powerful in deceiving a sentiment classifier than a direct synonym substitution attack on the sentiment classifier. This observation can be easily explained: the NMT attack has the potential to introduce words with a high target sentiment (positive/negative) in the output English sequence, whilst the *direct attack* on the output English sequence can only make substitutions with synonyms, limiting how positive/negative a sequence can be made. Hence, it can be concluded that an attack on the NMT system to change the sentiment of the output translation can be more powerful than an equivalent direct attack on the sentiment classifier.

3.3. Human Evaluation

In this work, to develop a practical method to generate adversarial examples, Section 2 introduced two proxy measures:

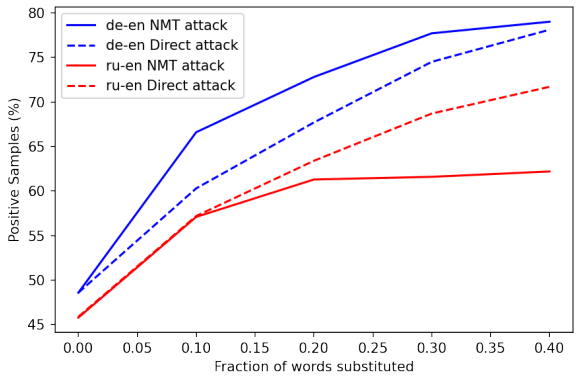
1. Four different measures and appropriate constraints were introduced to ensure that an adversarial example, $\tilde{x}$, is difficult to detect/imperceptible.
2. Equation 2 uses $\phi()$ as a proxy function to mimic human perception of the output sequence. The perception of interest in this work is *sentiment*.

The purpose of human evaluation experiments is to use a small, random sample of the data to verify the appropriateness of the automatic proxy measures/functions used in this work. Two sets of human evaluation experiments are conducted. In both sets of experiments, adversarial examples are generated using the systems, datasets and settings described in these experiments.

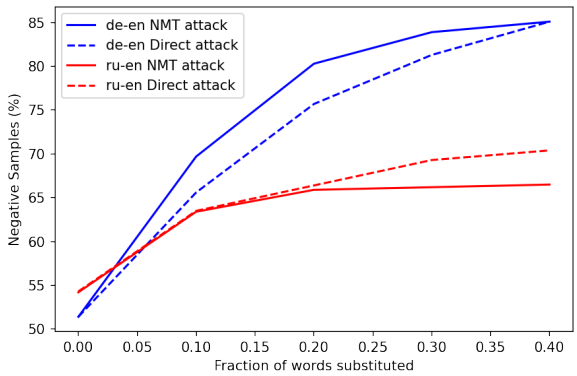
3.3.1. Adversarial Example Detectability

The aim of this human evaluation experiment is to assess the validity of the constraints (visual, semantic and perplexity) used to define an adversarial example. 50 original and 50 adversarial examples are selected randomly (and shuffled) for each translation model (ru-en, de-en, en-ru and en-de) and then human annotators [9] are asked to label each sample as *adversarial* or *authentic*. Table 4 gives the annotator

[9]Three volunteers carried out the human annotation in this work. For source languages German and Russian, the sentences

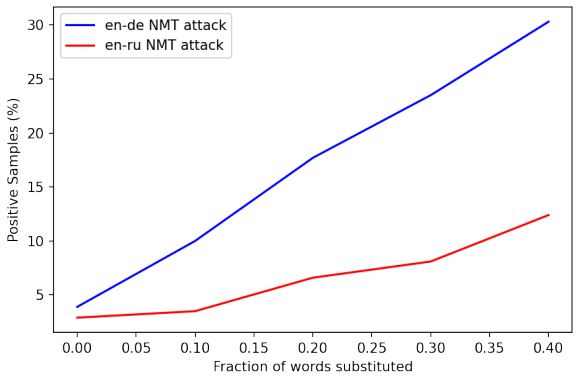


(a) Positive Sentiment

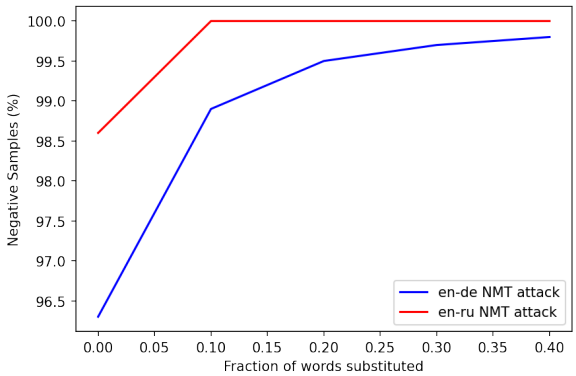


(b) Negative Sentiment

Figure 1. Perception adversarial attack on NMT systems to increase positive/negative sentiment. NMT attack: sentiment of predicted text with attack on input text. Direct attack: sentiment of predicted text with adversarial attack directly on predicted text.



(a) Positive Sentiment



(b) Negative Sentiment

Figure 2. Perception adversarial attack on NMT systems (with English as source language) to increase positive/negative sentiment.

accuracies for each task. The annotator accuracy is close to random (50%), suggesting that the adversarial examples are difficult to distinguish from authentic examples.

Model	Accuracy (%)
en-ru	54
en-de	52
de-en	49
ru-en	52

Table 4. Annotator accuracy for identifying adversarial/authentic examples

One further constraint of the adversarial attacks is that the sentiment of the source sequence should not change (Equation 5). Although this was enforced through a sentiment classifier applied on the source sequence and limiting the substitutions to synonyms, it is useful to run human evaluation experiments to ensure that the sentiment of adversarial samples truly did not change. For each translation task, 50 random pairs of adversarial and original examples were selected, and human annotators were required to state the sentiment of the original and adversarial examples separately. For both tasks with English as the source sequence, in all examples there was found to be no change in sentiment. For Russian as the source sequence, there were found to be 98% samples with no change and for German 96% sequences with no change in sentiment. This demonstrates that the adversarial attack method is almost perfectly satisfying the constraint that there is no change in the source sequence sentiment.

are also translated back to English using Google Translate to try and more easily identify adversarial examples.

3.3.2. Translation Sentiment

The aim of the adversarial attacks in this work is to change the output sentiment of Neural Machine Translation (NMT) systems. The sentiment of the output sequences is measured using proxy sentiment classifiers. It is necessary to ensure that the adversarial attack *only* attacks the NMT system and not the down-stream sentiment classifier; i.e. human evaluation is necessary to verify that the sentiment of a

translation, with an adversarial example at the input, as per the proxy sentiment classifiers, aligns with human perception of sentiment - we do not want a setting where an adversarial attack at the NMT input generates an output sequence with no change in human perception, but changes the sentiment classifier's output (an attack on the sentiment classifier). Two human evaluation experiment settings are considered: 1) sentiment of original translations - this is to give a reference proxy classifier performance; 2) sentiment of adversarial translations. As with the previous human evaluation experiment, 50 *positive* translations and 50 *negative* translations (as per the automatic sentiment classifiers) are randomly sampled per translation model for each setting. Human annotators are required to label each translation example as positive/negative. Table 5 gives the accuracy of these annotations with respect to the proxy sentiment classifiers. In the original setting, the classifier accuracy is around 93% for both NMT model outputs. There is only a small drop in accuracy of the sentiment classifiers in the adversarial setting, meaning there is still a strong agreement between human perception of sentiment and the classifiers' predicted sentiments for these translations [10]. Hence, it is argued that the adversarial attacks on the NMT systems are pre-dominantly attacking the NMT systems, as opposed to generating outputs that attack the sentiment classifiers.

Model	Original (%)	Adversarial (%)
de-en	93	89
ru-en	92	87

Table 5. Annotator consistency/accuracy with respect to the proxy sentiment classifier for sentiment prediction of translations of original and adversarial examples (adversarial attack is at the input of the NMT system- not the translation). In both settings the proxy classifier is consistent with sentiment perception of human annotations.

As the results from the human evaluation experiments correlate to a large extent with the automatic proxy measures/functions in this work, it is appropriate to use the automatic, human-free methods to generate adversarial samples at scale (i.e. using the entire dataset).

4. Conclusions

State-of-the-art sequence-to-sequence systems, including Neural Machine Translation (NMT) systems, have demonstrated vulnerability to adversarial attacks. These attacks involve subtle modifications to the input sequence that result in substantial changes in the output sequence, all while being imperceptible to human observers. While existing research has explored adversarial attack methods for NMT systems, focusing on the insertion of target phrases in the output sequences, this study contends that such attacks fail to encompass the entire spectrum of adversarial possibilities.

It is argued that adversaries seeking to manipulate NMT systems may aim to alter the perception of the output translation rather than merely inserting specific phrases. This research sheds light on the ease with which the sentiment perception of NMT system translations can be manipulated. By making minor modifications to the source language text without altering the underlying sentiment, the perception of sentiment in the output translation can be significantly distorted. This finding highlights the importance of considering the susceptibility of NMT systems to perception-based adversarial attacks.

Moving forward, further investigations will be undertaken to explore the robustness of other sequence-to-sequence systems against perception-based adversarial attacks. By broadening the scope of research in this area, a more comprehensive understanding of the vulnerabilities and potential defense mechanisms can be developed, ultimately contributing to the advancement of secure and reliable sequence-to-sequence systems.

5. Limitations

This work broadened the concept of adversarial attacks on sequence to sequence systems, where the aim is to change a human's *perception* of the generated output sequence. As a demonstration of this concept, this work explored changing the sentiment perception of the output sequence. It would be useful to also have consideration of other forms of perception of interest, e.g. *fluency* of the generated output sequence for language assessment tasks. This work presents results for four language pairs (de-en, ru-en, en-de, en-ru), and hence the applicability of this work to a language with a vastly different morphology may be limited without explicit experimentation. Finally, the method proposed in this work is applicable to any form of sequence-to-sequence task. Therefore, it would be useful to extend the experiments to tasks beyond machine translation, such as summarisation, question generation, question answering and even grammatical error correction.

Acknowledgements

This paper reports on research supported by Cambridge University Press & Assessment (CUP&A), a department of The Chancellor, Masters, and Scholars of the University of Cambridge.

[10] Only the ru-en and de-en models are considered in this experiment, as the other models/datasets contain a significant sentiment bias as is visible in Figure 2.

References

Blohm, M., Jagfeld, G., Sood, E., Yu, X., and Vu, N. T. Comparing attention-based convolutional and recurrent neural networks: Success and limitations in machine reading comprehension. *CoRR*, abs/1808.08744, 2018. URL http://arxiv.org/abs/1808.08744.

Carrara, F., Becarelli, R., Caldelli, R., Falchi, F., and Amato, G. Adversarial examples detection in features distance spaces. In Leal-Taixé, L. and Roth, S. (eds.), *Computer Vision – ECCV 2018 Workshops*, pp. 313–327, Cham, 2019. Springer International Publishing. ISBN 978-3-030-11012-3.

Cheng, M., Yi, J., Zhang, H., Chen, P., and Hsieh, C. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. *CoRR*, abs/1803.01128, 2018. URL http://arxiv.org/abs/1803.01128.

Ebrahimi, J., Lowd, D., and Dou, D. On adversarial examples for character-level neural machine translation. *CoRR*, abs/1806.09030, 2018. URL http://arxiv.org/abs/1806.09030.

Fellbaum, C. *WordNet: An electronic lexical database*. MIT Press, 1998.

Foundation, W. Acl 2019 fourth conference on machine translation (wmt19), shared task: Machine translation of news. URL http://www.statmt.org/wmt19/translation-task.html.

Goodfellow, I., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL http://arxiv.org/abs/1412.6572.

Iyyer, M., Wieting, J., Gimpel, K., and Zettlemoyer, L. 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)*, pp. 1875–1885, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1170. URL https://aclanthology.org/N18-1170.

Jia, R. and Liang, P. Adversarial examples for evaluating reading comprehension systems. *CoRR*, abs/1707.07328, 2017. URL http://arxiv.org/abs/1707.07328.

Li, J., Ji, S., Du, T., Li, B., and Wang, T. Textbugger: Generating adversarial text against real-world applications. *CoRR*, abs/1812.05271, 2018. URL http://arxiv.org/abs/1812.05271.

Michel, P., Li, X., Neubig, G., and Pino, J. M. On evaluation of adversarial perturbations for sequence-to-sequence models. *CoRR*, abs/1903.06620, 2019. URL http://arxiv.org/abs/1903.06620.

Neekhara, P., Hussain, S., Dubnov, S., and Koushanfar, F. Adversarial reprogramming of sequence classification neural networks. *CoRR*, abs/1809.01829, 2018. URL http://arxiv.org/abs/1809.01829.

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., and Edunov, S. Facebook fair's WMT19 news translation task submission. *CoRR*, abs/1907.06616, 2019. URL http://arxiv.org/abs/1907.06616.

odenet. Hdasprachtechnologie/odenet: Open german wordnet. URL https://github.com/hdaSprachtechnologie/odenet.

Post, M. A call for clarity in reporting BLEU scores. *CoRR*, abs/1804.08771, 2018. URL http://arxiv.org/abs/1804.08771.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language Models are Unsupervised Multitask Learners. 2019. URL https://openai.com/blog/better-language-models/.

Raina, V. and Gales, M. Residue-based natural language adversarial attack detection, 2022. URL https://arxiv.org/abs/2204.10192.

Raina, V., Gales, M. J., and Knill, K. M. Universal Adversarial Attacks on Spoken Language Assessment Systems. In *Proc. Interspeech 2020*, pp. 3855–3859, 2020. doi: 10.21437/Interspeech.2020-1890. URL http://dx.doi.org/10.21437/Interspeech.2020-1890.

Raina, V., Gales, M., and Lu, Y. Grammatical error correction systems for automated assessment: Are they susceptible to universal adversarial attacks? *Asia-Pacific Chapter of the Association for Computational Linguistics*, 2022. doi: https://doi.org/10.17863/CAM.88891.

Ren, S., Deng, Y., He, K., and Che, W. Generating natural language adversarial examples through probability weighted word saliency. In *ACL (1)*, pp. 1085–1097, 2019. URL https://doi.org/10.18653/v1/p19-1103.

Sakaguchi, K., Post, M., and Durme, B. V. Grammatical error correction with neural reinforcement learning. *CoRR*, abs/1707.00299, 2017. URL http://arxiv.org/abs/1707.00299.

Sun, M., Tang, F., Yi, J., Wang, F., and Zhou, J. Identify susceptible locations in medical records via adversarial attacks on deep predictive models. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery; Data Mining*, KDD '18, pp. 793–801, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3219909. URL https://doi.org/10.1145/3219819.3219909.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. 12 2013.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

wiki-ru wordnet. Wiki-ru-wordnet. URL https://pypi.org/project/wiki-ru-wordnet/.

Zhang, X., Zhang, J., Chen, Z., and He, K. 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)*. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.153. URL https://doi.org/10.18653%2Fv1%2F2021.acl-long.153.

Zhao, Z., Dua, D., and Singh, S. Generating natural adversarial examples. *CoRR*, abs/1710.11342, 2017. URL http://arxiv.org/abs/1710.11342.

Zou, W., Huang, S., Xie, J., Dai, X., and Chen, J. A reinforced generation of adversarial samples for neural machine translation. *CoRR*, abs/1911.03677, 2019. URL http://arxiv.org/abs/1911.03677.