Perception Adversarial Attacks on Neural Machine Translation Systems

Anonymous submission

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 large, 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 sequence-to-sequence tasks are explored from an output perception perspective. Thus the aim of an attack is to change the perception of the output sequence. For example, an adversary may want to make an output sequence have an exaggerated positive sentiment. In practice it is not possible 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 only small, imperceptible changes at the input sequences [1].

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 (Lin et al., 2014; Samanta and Mehta, 2017; Rosenberg et al., 2017; Huang et al., 2018; Papernot et al., 2016; Grosse et al., 2016; Sun et al., 2018; Cheng et al., 2018; Blohm et al., 2018; Neekhara et al., 2018; Jia and Liang, 2017; Niu and Bansal, 2018; Ribeiro et al., 2018; 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 miss-classifies (e.g. sentiment classification).

NMT systems, however, perform a sequence-to-sequence 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 adversarial attack needs to be modified for sequence-to-sequence tasks. Cheng et al. (2018) expands the adversarial attack definition for sequence-to-sequence models by introducing 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. Although this is a realistic setting for an adversarial attack, it does not capture attacks that seek to change the *perception* of the output sequence. An adversary may, for example, want to change the input text (in an imperceptible manner) such that the output text reads excessively 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*.

To the best of our knowledge, an attack on the perception of sequential outputs has not previously been examined. Thus, the main contribution of this work is the generalisation of the definition of adversarial attacks for sequence-to-sequence systems to include attacks that target the *perception* of the output. To demonstrate this form of attack, we perform experiments to change the sentiment of the output of NMT systems.

[1] Code is available at: *GitHub repository link will be provided after the anonymity period.*

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, $\hat{y}_{1:\hat{L}}$,

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

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,

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

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. To prevent easy detection of adversarial examples, it is necessary for the adversarial attack to satisfy an imperceptibility constraint, $\mathcal{G}()$, which again mimics human perception,

$$\mathcal{G}(x_{1:T}, \tilde{x}_{1:\tilde{T}}) \leq \epsilon, \tag{3}$$

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. This works uses a normalised variant of a Levenshtein, *edit-based* measurement (Li et al., 2018),

$$\mathcal{G}(x_{1:T}, \tilde{x}_{1:\tilde{T}}) = \frac{1}{T}\mathcal{L}(x_{1:T}, \tilde{x}_{1:\tilde{T}}). \tag{4}$$

The Levenshtein distance $\mathcal{L}()$ 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 work only examines word-level attacks, as these are considered more difficult to detect than character-level attacks (character level attacks can be easily detected using spelling and grammatical checks (Sakaguchi et al., 2017; Mays et al., 1991; Islam and Inkpen, 2009)). Specifically, this work restricts itself to an attack that substitutes $N = \epsilon T$ words (recall ϵ is the maximum fraction of edits permitted by the imperceptibility constraint). 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, $\tilde{x}_{1:\tilde{T}}$

$$\begin{aligned}\tilde{x}_{1:\tilde{T}} = x_1, \ldots, x_{n_1-1}, \tilde{x}_1, x_{n_1+1}, \ldots,\\ x_{n_N-1}, \tilde{x}_N, x_{n_N+1}, \ldots, x_T.\end{aligned} \tag{5}$$

It is challenging 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. To ensure little change in semantic content, only word synonyms are considered for the substitutions. In this work, the aim is to attack the perception of the output sequence (Equation 2). The mapping from input sequence, $x_{1:T}$ to perception score, $\phi()$, is non-differentiable, demanding a modified version of the saliency score for each word, $\mathcal{S}(x_t|x_{1:T})$, that measures the "sentiment saliency"

$$\begin{aligned}\mathcal{S}(x_t|x_{1:T}) = |\phi(\mathcal{F}_\theta(x_{1:t-1}, x_{t+1:T}))\\ - \phi(\mathcal{F}_\theta(x_{1:T}))|.\end{aligned} \tag{6}$$

3 Experiments

Experiments are performed using the NMT data from the WMT19 news translation task (Foundation). Results are presented for the Russian (ru) to English (en) and German (de) to English (en) tasks, where there are 2000 test examples. The best performing models, submitted by FAIR (Ng et al., 2019), are used as the baseline[2]. Table 1 gives the performance of these models on the WMT19 test set (respectively for each language), 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

Table 1: Model performances on WMT19 test sets

Each translation model is attacked using the saliency-based synonym substitution attack described in Equation 5, where the aim is to increase the *positivity* sentiment of the output English text sequence. The sentiment of the output

[2]NMT trained models available at: https://huggingface.co/facebook/wmt19-de-en and https://huggingface.co/facebook/wmt19-ru-en

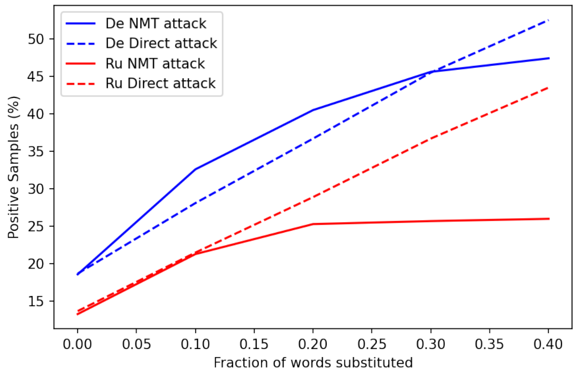

Figure 1: Perception adversarial attack on NMT systems to increase positive sentiment. NMT attack: sentiment of prediction sequence. Direct attack: sentiment of target reference text with adversarial attack directly on sentiment classifier.

sequence is measured using a standard, pre-trained (on 58M tweets) Roberta based English sentiment classifier [3]. Synonyms in the source Russian language are found using the `wiki-ru-wordnet` tool (wiki-ru wordnet), whilst synonyms for the source German language are given by the `OdeNet` tool (odenet). Examples of the attacks on the German NMT system are given in Table 2 [4].

Figure 1 shows the impact (curves *NMT attack*) of the adversarial attacks of increasing strength (fraction of words substituted), measured by the percentage of test samples classified as positive [5]. Both the Russian and German NMT systems observe more than a two-fold increase in the number of English translation samples classified as having a positive sentiment. For reference, the change in sentiment of the source Russian [6] and German [7] languages is also calculated. Both the German and Russian NMT systems, as expected, have a negligible increase in the positive sentiment of the source sequences. To be specific, when going from no attack to an attack strength of 40% of words substituted, the fraction of positive sentiment samples increases by 2% and 3%, for German and Russian source sequences respectively. This demonstrates that it is possible to have an imperceptible change at the input sequence (measured by sentiment and by the fraction of words substituted), that can cause a significant change in the perception of the output sequence.

Figure 1 gives one further curve for each NMT system: *direct attack*. Here, the same synonym substitution attack approach of Equation 5 is used to directly attack the output English sequence [8] to increase the positive sentiment score predicted by the English sentiment classifier. The substitutions are again limited to word synonyms. For the *ru-en* system, as would be expected, this direct attack of the sentiment classifier gives an upper-bound to the indirect NMT attack - an attack on the source language text is not expected to perform as well as an attack directly on the target language text. Note that this upper-bound direct attack is unrealistic for two reasons: 1) an adversary only has access to the source text; and 2) the direct attack on the English sequence is not imperceptible with respect to sentiment (for the *NMT attacks*, the Russian and German source texts had changed negligibly in their sentiments). However, the indirect *NMT attack* on the *de-en* NMT system is more powerful for up to 30% 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 positive sentiment in the output English sequence, whilst the *direct attack* on the output English sequence can only make substitutions with synonyms, limiting how positive 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.

All experiments in this section have used a simple metric to measure the success of adversarial attacks: the fraction of samples classified as positive, when using a max-class classifier. Table 3 presents equivalent results using instead the average (across the test dataset) predicted sentiment probabilities

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

[4] Examples of the attacks on the Russian NMT system are given in Table A.1

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

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

[7] German sentiment classifier: https://huggingface.co/oliverguhr/german-sentiment-bert

[8] This experiment used the reference English sequences from the WMT19 test sets.

	Original	Attacked
Source	Die Fans der Gunners, bei denen Granit Xhaka durchspielte, mussten sich gegen das Überraschungsteam aus der Grafschaft Hertfordshire allerdings bis in der 81. Minute auf eine Erfolgsmeldung gedulden.	Die Fans der Gunners, beiliegend denen Granit Xhaka durchspielte, mussten sich gegen das Überraschungsteam aus der Grafschaft Hertfordshire gewiss bis in der 81 . Minute unverstellt eine gute Kundmachung gedulden.
Prediction	However, the Gunners fans, with Granit Xhaka on the bench, had to wait until the 81st minute for news of their success against the surprise Hertfordshire team.	The Gunners fans, who were joined by Granit Xhaka, certainly had to endure a good display against the surprise Hertfordshire team until the 81st minute.
Sentiment	[0.21, 0.67, 0.12]	[0.01, 0.19, 0.80]
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	[0.35, 0.57, 0.08]	[0.01, 0.38, 0.61]

Table 2: Adversarial attack examples on de-en NMT system. Sentiment is: [negative, neutral, positive].

frac	Russian			German		
	Negative	Neutral	Positive	Negative	Neutral	Positive
ref	$0.223_{\pm 0.272}$	$0.600_{\pm 0.275}$	$0.178_{\pm 0.259}$	$0.221_{\pm 0.270}$	$0.556_{\pm 0.270}$	$0.223_{\pm 0.287}$
0	$0.224_{\pm 0.274}$	$0.603_{\pm 0.277}$	$0.173_{\pm 0.256}$	$0.223_{\pm 0.272}$	$0.556_{\pm 0.271}$	$0.219_{\pm 0.284}$
0.1	$0.180_{\pm 0.246}$	$0.566_{\pm 0.263}$	$0.257_{\pm 0.292}$	$0.132_{\pm 0.205}$	$0.530_{\pm 0.275}$	$0.338_{\pm 0.327}$
0.2	$0.162_{\pm 0.234}$	$0.548_{\pm 0.262}$	$0.290_{\pm 0.303}$	$0.101_{\pm 0.175}$	$0.491_{\pm 0.284}$	$0.408_{\pm 0.342}$
0.3	$0.160_{\pm 0.232}$	$0.546_{\pm 0.264}$	$0.294_{\pm 0.306}$	$0.085_{\pm 0.157}$	$0.466_{\pm 0.287}$	$0.447_{\pm 0.343}$
0.4	$0.158_{\pm 0.231}$	$0.545_{\pm 0.262}$	$0.297_{\pm 0.305}$	$0.080_{\pm 0.151}$	$0.456_{\pm 0.287}$	$0.464_{\pm 0.344}$

Table 3: Average (over test dataset) sentiment probability (with frac percentages of words substituted) of ru/de-en NMT system's predicted sequence. ref is the reference target English sequence.

(for each of the negative, neutral and sentiment classes). The results in this table also indicate the standard deviation over the test dataset. The trends visible for the positive class in Table 3 are identical to the trends identified so far in this section - the average positive probability increases significantly with the adversarial attack. When considering the negative and neutral classes, it can be seen that for small attack strengths (frac=0.1), the average negative probability decreases dramatically, whilst the neutral class probability surprisingly increases. This is revealing of an ordering of the sentiment classes: the adversarial attack converts negative prediction sequences into more neutral prediction sequences, which in turn are transformed into more positive prediction sequences.

4 Conclusions

Best performing sequence-to-sequence systems, such as Neural Machine Translation systems, are dominated by deep learning based architectures. Like other deep learning systems, sequence-to-sequence systems are also vulnerable to adversarial attacks. An adversary can make a small, imperceptible change to the input sequence that causes a significant change in the output sequence. For NMT systems, existing works in literature propose adversarial attack methods that are designed to insert target phrases in the output sequences. This work argues that this form of attack is not encompassing of all styles of adversarial attacks. Specifically, an adversary may attempt to change the *perception* of the output translation, as opposed to inserting some target phrase. This work shows that the perception of sentiment, as measured by a standard sentiment classifier, of the output translation of NMT systems can be easily attacked, where only small changes are made to the source language text. Future work will explore robustness of sequence-to-sequence systems to perception adversarial attacks.

References

Matthias Blohm, Glorianna Jagfeld, Ekta Sood, Xiang Yu, and Ngoc Thang Vu. 2018. Comparing attention-based convolutional and recurrent neural networks: Success and limitations in machine reading comprehension. *CoRR*, abs/1808.08744.

Fabio Carrara, Rudy Becarelli, Roberto Caldelli, Fabrizio Falchi, and Giuseppe Amato. 2019. Adversarial examples detection in features distance spaces. In *Computer Vision – ECCV 2018 Workshops*, pages 313–327, Cham. Springer International Publishing.

Minhao Cheng, Jinfeng Yi, Huan Zhang, Pin-Yu Chen, and Cho-Jui Hsieh. 2018. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. *CoRR*, abs/1803.01128.

Javid Ebrahimi, Daniel Lowd, and Dejing Dou. 2018. On adversarial examples for character-level neural machine translation. *CoRR*, abs/1806.09030.

Wikimedia Foundation. Acl 2019 fourth conference on machine translation (wmt19), shared task: Machine translation of news.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

Kathrin Grosse, Nicolas Papernot, Praveen Manoharan, Michael Backes, and Patrick D. McDaniel. 2016. Adversarial perturbations against deep neural networks for malware classification. *CoRR*, abs/1606.04435.

Alex Huang, Abdullah Al-Dujaili, Erik Hemberg, and Una-May O'Reilly. 2018. Adversarial deep learning for robust detection of binary encoded malware. *CoRR*, abs/1801.02950.

Aminul Islam and Diana Inkpen. 2009. Real-word spelling correction using google web it 3-grams. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, page 1241–1249, USA. Association for Computational Linguistics.

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)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *CoRR*, abs/1707.07328.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2018. Textbugger: Generating adversarial text against real-world applications. *CoRR*, abs/1812.05271.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Eric Mays, Fred J. Damerau, and Robert L. Mercer. 1991. Context based spelling correction. *Information Processing Management*, 27(5):517–522.

Paarth Neekhara, Shehzeen Hussain, Shlomo Dubnov, and Farinaz Koushanfar. 2018. Adversarial reprogramming of sequence classification neural networks. *CoRR*, abs/1809.01829.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's WMT19 news translation task submission. *CoRR*, abs/1907.06616.

Tong Niu and Mohit Bansal. 2018. Adversarial over-sensitivity and over-stability strategies for dialogue models. *CoRR*, abs/1809.02079.

odenet. Hdasprachtechnologie/odenet: Open german wordnet.

Nicolas Papernot, Patrick D. McDaniel, Ananthram Swami, and Richard E. Harang. 2016. Crafting adversarial input sequences for recurrent neural networks. *CoRR*, abs/1604.08275.

Matt Post. 2018. A call for clarity in reporting BLEU scores. *CoRR*, abs/1804.08771.

Vyas Raina, Mark J.F. Gales, and Kate M. Knill. 2020. Universal Adversarial Attacks on Spoken Language Assessment Systems. In *Proc. Interspeech 2020*, pages 3855–3859.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *ACL (1)*, pages 1085–1097.

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, Melbourne, Australia. Association for Computational Linguistics.

Ishai Rosenberg, Asaf Shabtai, Lior Rokach, and Yuval Elovici. 2017. Generic black-box end-to-end attack against rnns and other API calls based malware classifiers. *CoRR*, abs/1707.05970.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2017. Grammatical error correction with neural reinforcement learning. *CoRR*, abs/1707.00299.

Suranjana Samanta and Sameep Mehta. 2017. Towards crafting text adversarial samples. *CoRR*, abs/1707.02812.

Mengying Sun, Fengyi Tang, Jinfeng Yi, Fei Wang, and Jiayu Zhou. 2018. 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 amp; Data Mining*, KDD '18, page 793–801, New York, NY, USA. Association for Computing Machinery.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks.

wiki-ru wordnet. Wiki-ru-wordnet.

Xinze Zhang, Junzhe Zhang, Zhenhua Chen, and Kun He. 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)*. Association for Computational Linguistics.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2017. Generating natural adversarial examples. *CoRR*, abs/1710.11342.

Wei Zou, Shujian Huang, Jun Xie, Xinyu Dai, and Jiajun Chen. 2019. A reinforced generation of adversarial samples for neural machine translation. *CoRR*, abs/1911.03677.

Appendix A

A.1 Results

	Original	Attacked
Source	Он также не исключил, что реальные цифры призванных в армию украинцев могут быть увеличены в случае необходимости.	Он также не исключил, что реальные цифры призванных в армию украинцев могут быть увеличены во благо случае необходимости.
Prediction	He also did not rule out that the real number of Ukrainians drafted into the army could be increased if necessary.	He also did not rule out that the real number of Ukrainians drafted into the army could be increased for good if necessary.
Sentiment	[0.11, 0.84, 0.04]	[0.03, 0.66, 0.30]
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	[0.04, 0.76, 0.21]	[0.01, 0.50, 0.48]

Таблица A.1: Adversarial attack examples on ru-en NMT system. Sentiment is: [negative, neutral, positive].

frac	Russian			German		
	Negative	Neutral	Positive	Negative	Neutral	Positive
0.0	$0.223_{\pm 0.272}$	$0.600_{\pm 0.275}$	$0.178_{\pm 0.259}$	$0.221_{\pm 0.270}$	$0.556_{\pm 0.270}$	$0.223_{\pm 0.287}$
0.1	$0.116_{\pm 0.180}$	$0.625_{\pm 0.266}$	$0.258_{\pm 0.292}$	$0.109_{\pm 0.175}$	$0.576_{\pm 0.272}$	$0.315_{\pm 0.312}$
0.2	$0.079_{\pm 0.135}$	$0.591_{\pm 0.274}$	$0.330_{\pm 0.310}$	$0.072_{\pm 0.129}$	$0.538_{\pm 0.281}$	$0.390_{\pm 0.323}$
0.3	$0.055_{\pm 0.102}$	$0.548_{\pm 0.280}$	$0.397_{\pm 0.315}$	$0.050_{\pm 0.099}$	$0.493_{\pm 0.287}$	$0.457_{\pm 0.325}$
0.4	$0.042_{\pm 0.082}$	$0.509_{\pm 0.281}$	$0.449_{\pm 0.313}$	$0.037_{\pm 0.080}$	$0.453_{\pm 0.284}$	$0.511_{\pm 0.317}$

Table A.2: Average (over test dataset) sentiment probability (with frac percentages of words substituted in synonym substitution adversarial attack) of ru/de-en NMT system's **target** sequence (*direct attack* on Roberta based sentiment classifier).

A.2 Limitations

The perception adversarial attack experiments in this work focus solely on NMT systems, as opposed to considering a range of sequence to sequence systems. A further limitation is that the perception of output sequences is measured using proxy deep models as opposed a direct human evaluation. However, human evaluations are expensive and impractical for large scale experiments.

A.3 Risks and Ethics

Experiments in this work reveal methods by which an adversary can deceive state of the art, deployed Neural Machine Translation systems. However, these forms of attacks are in their infancy and therefore it is not considered a realistic threat for real-world applications.