Записки научных
семинаров ПОМИ
Том 530, 2023 г.

A. Chertkov, O. Tsymboi, M. Pautov, I. Oseledets

TRANSLATE YOUR GIBBERISH: BLACK-BOX ADVERSARIAL ATTACK ON MACHINE TRANSLATION SYSTEMS

Abstract. Neural networks are deployed widely in natural language processing tasks on the industrial scale, and perhaps most often they are used as compounds of automatic machine translation systems. In this work, we present a simple approach to fool state of the art machine translation tools in the task of translation from Russian to English and vice versa. Using a novel black-box gradient-free tensor-based optimizer, we show that many online translation tools, such as Google, DeepL, and Yandex, may both produce wrong or offensive translations for nonsensical adversarial input queries and refuse to translate seemingly benign input phrases. This vulnerability may interfere with understanding a new language and simply worsen the user's experience while using machine translation systems, and, hence, additional improvements of these tools are required to establish better translation.

§1. Introduction

Adversarial perturbations are carefully crafted modifications of the input that are imperceptible for humans but force a machine learning model to perform poorly. Initially discovered in the domain of computer vision [16, 27], where imperceptibility is attained by restricting the norm of additive perturbation, they were later extended to the natural language processing (NLP). Since the nature of language is discrete, the imperceptibility in NLP is attained either on the character level [12,14], where only few characters in a word are subject to change, or on the word level [4, 6], where the words are allowed to be replaced only by semantically similar words (e.g., by synonyms).

However, machine translation (MT) systems are known to be vulnerable to adversarial examples with relaxed imperceptibility [5]. More than that, apart from sensitivity to imperceptible adversarial examples, MT may both

Key words and phrases: natural language processing, machine translation, adversarial attack, black-box optimization .

produce meaningful translations for nonsensical gibberish input queries and refuse to translate seemingly benign input phrases. This unpredictable behavior may not only interfere with understanding a new language but also may lead to serious problems (e.g., several years ago Facebook's MT system mistranslated an Arabic phrase meaning "good morning" as "attack them" which led to a wrongful arrest [3, 13]). Hence, understanding the unpredictable behavior of these systems is an essential step for improving the robustness of machine translation and, as a result, for preventing such incidents.

In this work, we investigate the stability and behavior of MT systems for inputs with low likelihood. We consider three major well-known online translators DeepL Google, and Yandex, and set the task of automatically finding an input in Russian representing an arbitrary set of letters of a given length (not a word), which, however, leads to a meaningful translation into English (a word or set of words). We formulate it as a problem of maximizing the difference between the perplexity [25] of the translation and the source text, and we apply GPT-2 [22] to define the perplexity of the input and output sequences. For a search of the best combination of input symbols we use the new optimization method PROTES[1] [2], which is based on the low-rank tensor train (TT) decomposition [21] and can efficiently perform gradient-free multivariate discrete optimization. For all three considered MT systems, we obtained a set of seven-letter inputs in Russian that are not words, which, however, lead to a translation representing a word or set of words in English. Hereafter, for the sake of brevity, we will refer to such inputs as *hallucinogens*. What is an intriguing, both manual and automatic combinations of the obtained hallucinogens, as it turned out, allows getting a variety of valid English phrases. Moreover, some of these phrases turn out to be examples of adversarial attacks (detected so far only for the DeepL translator). When trying to translate them back into Russian, the translator produces significantly incorrect results (garbage word combinations or even a blank translation string). To summarize, our contributions are the following:

- we develop a new black-box optimization method for the automatic generation of low-likelihood input sequences ("hallucinogens") with high translation likelihood for MT systems based on the perplexity estimation of the input and output sequences;

[1] We use the code from `https://github.com/anabatsh/PROTES`.

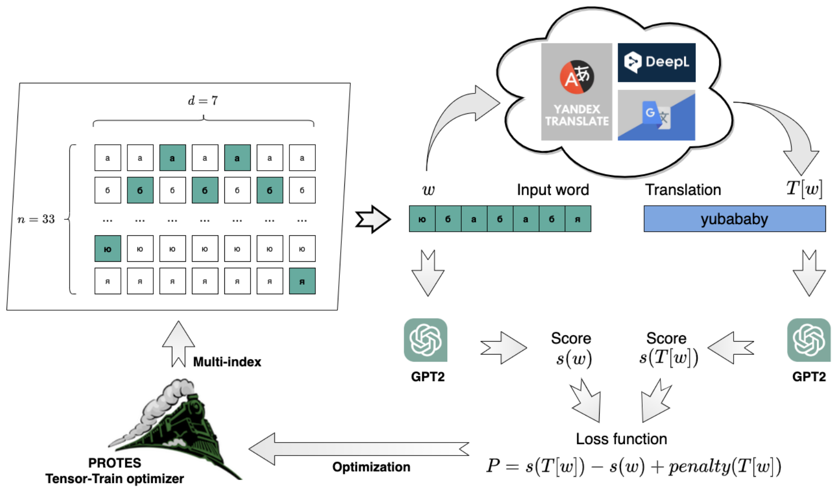

Figure 1. Proposed approach for searching for the "hallucinogens".

- we demonstrate that it is possible to use this approach for black-box adversarial attacks on MT systems since the corresponding translation results for a set (phrase) of hallucinogens often correspond to the "instability points" of the system and lead to invalid backward translation;
- we apply[2] the proposed approach for major online translators DeepL, Google, and Yandex, find an extensive set of hallucinogens and their combinations for all three translators, and demonstrate the possibility of an adversarial attack on the DeepL system.

§2. Method

Our approach is presented in Figure 1 and is based on the idea of searching for d-letter combinations $w = (w_1, w_2, \ldots, w_d)$ in the source language that are the least similar to existing words (gibberish or "hallucinogens") but are, however, correctly translatable into the target language as $\mathsf{T}[w]$. Without loss of generality, we have chosen Russian as the source language

[2]The program code and all results with the supporting screenshots are available in our public repository `https://github.com/AndreiChertkov/TranFighterPro`.

(it has $n = 33$ letters of the alphabet), English language as the target language (it has $n_t = 26$ letters of the alphabet), and $d = 7$.

To assess the quality (score) of a word or phrase, we use perplexity [25]

$$\mathsf{s}(w) = \mathsf{exp}\left[-\frac{1}{d}\sum_{i=1}^{d}\log p_\theta(w_i|w_{<i})\right], \tag{1}$$

where $p_\theta(w_i|w_{<i})$ is the log-likelihood of the i-th token conditioned on the preceding tokens according to the pre-trained GPT-2 model. It can be thought of as an evaluation of the model's ability to predict among the set of specified tokens in a corpus. The value $\mathsf{s}(w)$ is non-negative, for the most common words it is close to zero, and for the gibberish, it is expected to be a large positive number.

To maximize the difference between the perplexity of the translation $\mathsf{T}[w]$ and the source text w we introduce the following loss function:

$$P(w) = \mathsf{s}(\mathsf{T}[w]) - \mathsf{s}(w) + \mathsf{penalty}(\mathsf{T}[w]), \tag{2}$$

where $\mathsf{penalty}(\mathsf{T}[w])$ is a penalty term, which is equal to a large positive number for the case when the translation is too short (less than 5 characters) or contains stop characters (various non-letter characters); otherwise it is zero.

We search for the minimum of (2) in terms of the discrete optimization problem for an implicitly given d-dimensional array $\mathcal{P} \in \mathbb{R}^{n\times n\times\ldots\times n}$:

$$\mathcal{P}[i_1, i_2, \ldots, i_d] = P(w), \quad w = (A[i_1], A[i_2], \ldots, A[i_d]), \tag{3}$$

where $[i_1, i_2, \ldots, i_d]$ is a multi-index, A is the alphabet, and $A[i_k]$ is the i_k-th symbol of the alphabet. For example, as shown in Figure 1, for the multi-index [32, 2, 1, 2, 1, 2, 33] we get the word w "юбабабя" in Russian.

To find the "hallucinogen" $\hat{w}$ which minimizes the loss function (2), we use the global optimization method PROTES. It is based on the low-rank tensor train (TT) decomposition [8–10, 21, 26], which allows bypassing the curse of dimensionality problem[3]. The method operates with a multidimensional discrete probability distribution in the TT-format, followed by efficient sampling from it and updating its parameters by stochastic gradient ascent to approximate the minimum or maximum in a better way.

[3]The complexity of algorithms in the TT-format (e. g., element-wise addition, multiplication, solution of linear systems, convolution, integration, etc.) turns out to be polynomial in dimension and mode size, and it makes TT-decomposition extremely popular in a wide range of applications, including computational mathematics and machine learning.

Table 1. Top-33 generated hallucinogens for the DeepL translator.

Text	Translation	Loss	Text	Translation	Loss	Text	Translation	Loss
быелръъ	formerly	-42.52	оощвишн	Promotion	-26.86	гзйкщчж	gzcjcj	-23.04
пдлешйщ	Synopsis:	-39.47	ощуъиъв	Feelings	-25.08	ъоэсйьл	Yoesyl	-22.33
бысёъгч	Quickly	-38.53	гбьъьиэ	gbjie	-24.08	мжвлвфж	mjvlvfj	-22.0
чтьёиэе	READ MORE	-37.2	рыьдяно	snarky	-24.07	ктлтксь	ktltx	-21.61
щосющйе	Synopsis:	-34.84	жьрэиэф	zhreif	-23.64	фйвьжиы	fyvji	-21.38
бынщийя	former	-34.84	жцчыщцй	Žučičky	-23.64	жаьйщсч	zhayshch	-21.25
зсзгвлэ	ssgvle	-30.42	чёхёшьч	What the fuck	-23.49	ккзёйьи	kkzoyi	-20.78
бгаьъэы	bgaiy	-30.12	зжнмкьъ	zznnmkj	-23.37	бфзскйт	bfzskyt	-20.66
дачэщйч	Dachshund	-27.67	гмххъьн	gmhxjn	-23.21	ыьбэъхс	yybexx	-20.47
бреощее	Breaking	-27.5	жьрцэъо	Jrceo	-23.19	ъйлбмфь	ylbmfj	-20.27
бжкльлш	bjklsh	-27.21	бёацсжю	boatsjue	-23.15	чъръпьм	chirp	-20.23

We save the request history of the optimization method and, at the end of its run, we form a set of hallucinogens $\hat{w}^{(1)}, \hat{w}^{(2)}, \ldots, \hat{w}^{(m)}$ (m here is the number of requests to a translator, i.e., the computational budget), ordered by the value of the loss function.

It is worth mentioning that the described method does not generate adversarial examples per se (i.e., it does not force mistranslation) but produces examples (hallucinogens) that are translatable when they should not be. However, it turns out to be an interesting empirical fact that combinations of hallucinogens also lead to the emergence of translation artifacts, while, as we will show below, these artifacts can turn out to be long meaningful phrases in the target language.

Accordingly, in the second stage we repeat the described optimization process, composing phrases of $d^{(2)}$ hallucinogens. As possible candidates, we select $n^{(2)}$ ($n^{(2)} \leq m$) top hallucinogens $\hat{w}^{(1)}, \hat{w}^{(2)}, \ldots, \hat{w}^{(n^{(2)})}$ from the results of the first stage. Without loss of generality, we have chosen $d^{(2)} = 7$ and $n^{(2)} = 33$, i.e., the same values as in the first stage. In this case, we use the loss function (2) without the second term, i.e., we do not maximize the perplexity of the input text, since it is already composed of hallucinogens. Note that we can repeat this process an arbitrary number of times, getting longer and longer "phrases" from the hallucinogens.

§3. Experiments

We consider three well-known online translators—DeepL, Google, and Yandex—and search for hallucinogens following the scheme presented in the previous section. For each translator, we limit the optimizer budget

Table 2. Top-33 generated hallucinogens for the Google translator.

Text	Translation	Loss	Text	Translation	Loss	Text	Translation	Loss
ъувщжёь	Knight	-50.18	штшнлхж	Stitch	-35.53	ьокнёйф	Continuity	-30.15
бйввкшя	Former	-48.27	гяшрьнп	Gagarin	-33.98	ъфъыхлч	Kommersant	-30.1
дщижщяп	Building	-45.13	здкънсп	health	-33.39	птйдфдц	PTDDC	-30.09
мощыъпз	Power	-43.64	ъыллщьн	Kommersant	-32.24	йтдкцяе	induction	-29.54
ъыьгрвх	Kommersant	-43.38	ътшлшэь	Kommersant	-32.0	уясъцёь	understanding	-29.29
пёвюмыц	first	-41.73	быошийя	To be	-31.81	зсзгвлэ	ZSZGLE	-29.28
ъёефнся	Currently	-41.19	доцшлны	Associated	-31.69	ъфоъкцж	Kommersant	-29.01
ъжлхчлы	Kommersant	-37.32	пщмёжны	They are	-31.62	жхнаеыь	grunts	-28.97
ъоэсйьл	Kommersant	-37.21	ъухвмгс	Kommersant	-31.38	ъфкщтнэ	Kommersant	-28.68
вытёщдч	priest	-37.05	ъбывзлц	Kommersant	-30.8	ъныуазу	Kommersant	-28.47
бщагчёщ	Passing	-36.29	бяёщжии	beads	-30.24	гфоаьын	fifajn	-28.38

Table 3. Top-33 generated hallucinogens for the Yandex translator.

Text	Translation	Loss	Text	Translation	Loss	Text	Translation	Loss
здблоьп	hello	-42.87	кмтсгфк	kmtsgfc	-27.48	иьллтёу	illteu	-24.03
ьвднэйу	Today	-42.15	иощсцйм	ioschcym	-27.08	щаафечу	right now	-23.68
онуьлйц	online	-40.44	нзеъёаь	nzeea	-26.32	ъяляужь	for the service	-23.41
смэёыюш	see also	-35.26	бмъчкьь	bmchk	-26.1	нмьрщшт	nmrsht	-23.33
ысвщёы	and more	-34.94	ъоэсйьм	yoesm	-25.67	оэеыъьё	oeeye	-23.16
схисеъм	scheme	-32.76	ъыклщьн	kommersant	-25.56	йьаёьеб	yaeeb	-23.1
мощыъпз	The power of the	-31.2	бьвтюья	byuya	-25.49	флжсйид	fljsyid	-22.72
кццжйхк	kccjhk	-30.76	иьеьрёъ	iyere	-25.48	пёыэулм	peeulm	-22.67
ътшмщэь	kommersant	-30.54	ущйинъу	pinyin	-25.22	бдлпроь	bdlpro	-22.59
ъубщжёь	kommersant	-27.58	шэьдкйя	shadkya	-24.49	доцшлмь	assoc .	-22.56
ььгрвх	ygrvh	-27.56	ощуъиъв	feeling	-24.03	ъныуазу	kommersant	-22.53

Table 4. Sample generated combinations of hallucinogens for the DeepL translator.

Text	Translation
жьрцэъо жьрцэъо ощуъиъв ъйлбмфь чтьёиэе ъйлбмфь зжнм-кьъ	Greetings from the Greetings Department of the Ministry of Foreign Affairs
быншийя бгаьъэы ъоэсйьл чёхёшьч мжвлвфж рыьдяно гзйкщчж	The formerly bogeyman is the one who is the most important person in the world.
бреощее бысёъгч жаьйщсч жьрэи-эф зсзгвлэ пдлешйщ оощвишн	The main reason for this is that we have a lot of time and effort to get to the bottom of this

Table 5. Sample generated combinations of hallucinogens for the Google translator.

Text	Translation
уясъцёь ъыллщьн пщмёжны ъныуазу йтдкцяе бщагчёщ ъёефнся	understanding of the bang
быошийя ъёефнся ъбывзлщ ъжлхчлы быошийя йтдкцяе пёвюмыщ	I would have been the bungles of Kommersant Kommersant
вытёщдч доцшлны ъувщжёь бйввкшя пщмёжны ъыллщьн бяёщжии	The priests of the Associate Professor Kommersant

Table 6. Sample generated combinations of hallucinogens for the Yandex translator.

Text	Translation
мощыъпз щаафечу йьаёьеб ощуъиъв нзеъёаь ощуъиъв иысвщёы	The power of the heart is now being felt by the heart of the heart .
ъяляужь иысвщёы иьллтёу оэеыъьё щаафечу мощыъпз ощуъиъв	I will be able to feel the power of the heart.
ощуъиъв доцшлмь ъныуазу онуьлйц ьвднэйу здблоьп ьвднэйу	I feel like I 'm on the right side of the right side of the right side of the right side of the right side of the right side of the right side

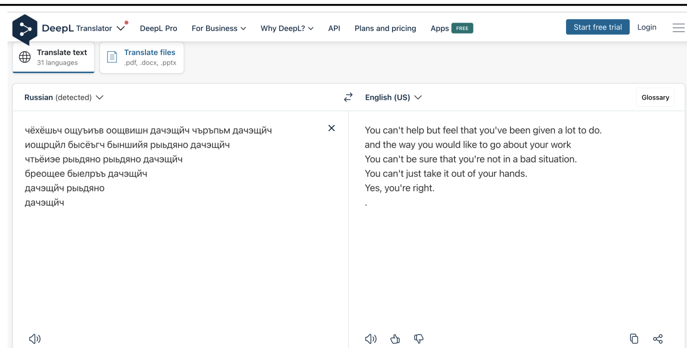

Figure 2. Composition of hallucinogens for the DeepL translator.

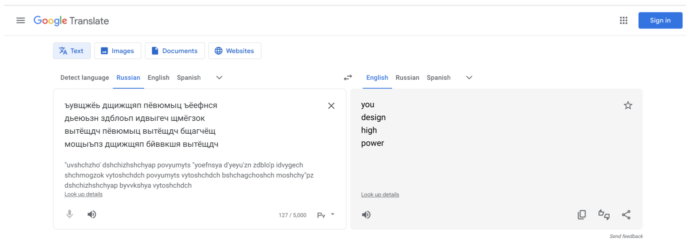

Figure 3. Composition of hallucinogens for the Google translator.

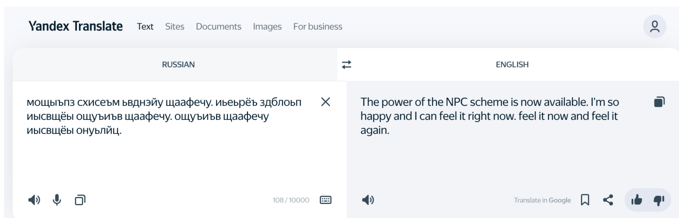

Figure 4. Composition of hallucinogens for the Yandex translator.

to $m = 1000$ translations and use the default values for the rest of the parameters.

Results[4] for DeepL, Google and Yandex are presented in Tables 1, 2 and 3, respectively. Note that using the found seven-letter hallucinogens in Russian, we can easily manually build funny examples for each of the translators, in which the junk text at the input is translated into correct text in English. We also refer to the related examples in Figures 2, 3 and 4.

Then we run the optimization process for phrases of top-7 hallucinogens from the first stage. The corresponding results are presented in Tables 4, 5 and 6. Note that optimization based on perplexity, in this case, yields phrases that are translatable into English but not always expressive enough (the complete list of phrases is presented in our repository). Therefore, in

[4]As of this writing, all of the results presented for DeepL and Yandex (and Figure 3 for Google) can be reproduced in a modern web browser. The results (see Tables 2 and 5) for Google translator were obtained with an older version of the browser (Chrome Canary 111.0.5555.0), which loads an older version of the translator, and are not fully reproducible in modern web browsers.

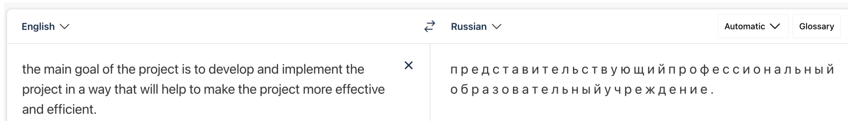

Figure 5. Backtranslation results for the attack text "фйвьжиы фйвьжиы пдлешйщ ккзёйьи гбьъьиэ жцчыщцй ктлткcь ыьбэъхс ъоэсйьл жьрцэъо мжвлвфж гзйкщчж жцчыщцй щосющйе ккзёйьи ккзёйьи фйвьжиы быншийя дачэщйч бысёъгч бёацсжю бысёъгч жцчыщцй жьрэиэф гмххъьн бёацсжю бгаьъэы чёхёшьч оощвишн бжкльлш бжкльлш щосющйе бгаьъэы дачэщйч ъоэсйьл пдлешйщ жцчыщцй жаьйщсч ъоэсйьл чёхёшьч бреощее ъйлбмфь бреощее бгаьъэы бжкльлш жьрэиэф ктлткcь ктлткcь бгаьъэы". The resulting Russian translation has the following meaning in English: "representative professional educational institution".

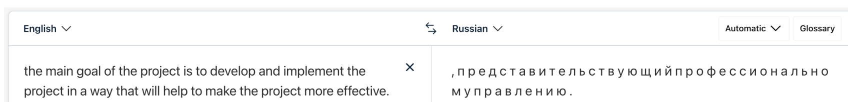

Figure 6. Backtranslation results for the attack text "бёацсжю бгаьъэы гзйкщчж фйвьжиы дачэщйч бысёъгч ккзёйьи ъоэсйьл гзйкщчж гбьъьиэ жьрэиэф зжнмкьъ бысёъгч бреощее жьрцэъо быелръъ жаьйщсч бреощее зжнмкьъ чъръпьм ъйлбмфь ккзёйьи гзйкщчж гбьъьиэ зсзгвлэ жьрцэъо гзйкщчж чтьёиэе бысёъгч жцчыщцй жьрэиэф гмххъьн бёацсжю бгаьъэы чёхёшьч чёхёшьч ктлткcь бысёъгч ъоэсйьл быелръъ чёхёшьч гмххъьн жьрэиэф бжкльлш зсзгвлэ жьрцэъо бысёъгч бысёъгч бжкльлш". The resulting Russian translation has the following meaning in English: "representing professional management".

the tables we report three hand-selected quite expressive results for each of the translators.

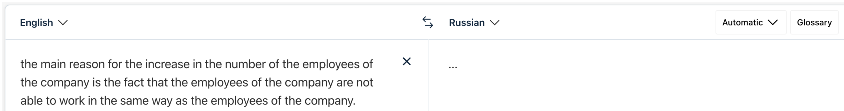

Figure 7. Backtranslation results for the attack text “рыьдяно рыьдяно фйвьжиы рыьдяно жьрэиэф щосющйе рыьдяно жцчыщцй фйвьжиы гбьъьиэ зсзгвлэ бгаъэы рыьдяно ккзёйьи ктлткcь бфзскйт щосющйе пдлешйщ мжвлвфж рыьдяно гзйкщчж зсзгвлэ гзйкщчж гзйкщчж гбьъьиэ оощвишн гзйкщчж чёхёшьч пдлешйщ жцчыщцй жаьйщсч ъоэсйьл чёхёшьч бреощее ъйлбмфь ктлткcь бфзскйт щосющйе пдлешйщ мжвлвфж рыьдяно гзйкщчж чъръпьм чъръпьм ъйлбмфь пдлешйщ быншийя ощуъиъв ыьбэъхс”. The resulting Russian translation is empty.

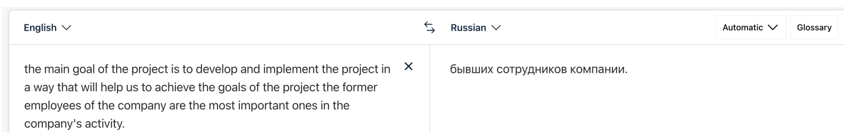

Figure 8. Backtranslation results for the attack text “бёацсжю бгаъэы гзйкщчж фйвьжиы дачэщйч бысёъгч ккзёйьи чёхёшьч ктлткcь бысёъгч ъоэсйьл быелръъ чёхёшьч гмххъьн ъоэсйьл ккзёйьи бжкльлш пдлешйщ рыьдяно жьрцэъо пдлешйщ бёацсжю зсзгвлэ бёацсжю чтьёиэе быншийя бжкльлш гзйкщчж чъръпьм чъръпьм ъйлбмфь пдлешйщ быншийя ощуъиъв ыьбэъхс бёацсжю бгаъэы бреощее зжнмкьъ жаьйщсч ктлткcь ккзёйьи оощвишн бжкльлш бжкльлш щосющйе бгаъэы дачэщйч ъоэсйьл”. The resulting Russian translation has the following meaning in English: “former employees of the company.”.

The same procedure is conducted for the DeepL translator with generation of longer sequences of hallucinogens. In this case, we use the top-33 phrases of 7 hallucinogens from the results of the second step, and, as before, compose their combinations of length 7 (that is, in this case we

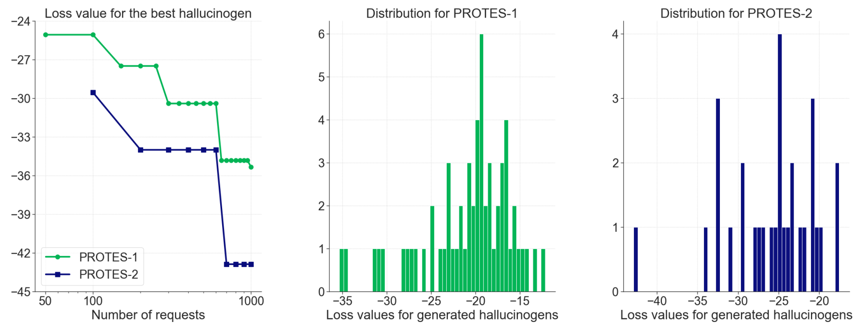

Figure 9. Dependence of the found optimum on the number of requests to the DeepL online translator (plot on the left) and the distribution of results (plots in the center and on the right) for two optimizer configurations.

are making a sequence of hallucinogens of length 49). As a result, an interesting fact was discovered: DeepL fails to translate back into Russian the resulting meaningful English phrases. In Figures 5–8 we report some related examples of adversarial attacks.

3.0.0.1. Parameters of the optimizer. In all experiments, we used the default set of parameters for PROTES (below we will call this configuration "PROTES-1"): $K = 50$ (the number of generated samples per iteration, i.e., the batch size), $k = 5$ (the number of selected candidates per iteration), $k_{gd} = 100$ (the number of gradient ascent steps), $\lambda = 10^{-4}$ (the gradient ascent learning rate), $R = 5$ (the TT-rank of the probability tensor), and we limit the number of requests to the translator at the value $m = 10^3$. To evaluate the influence of the choice of parameters on the final result, we also try the following configuration ("PROTES-2"): $K = 100$, $k = 10$, $k_{gd} = 1$, $\lambda = 0.05$, $R = 5$.

To compare the two sets of parameters[5], we consider the DeepL online translator, and in Table 7 we present the best-generated hallucinogens for each requested batch (that is, for every batch of 50 and 100 inputs for translation requested by the optimizer "PROTES-1" and "PROTES-2", respectively). In Figure 9 we present the dependence of the found optimum (i.e., the value of the loss function) on the number of requests and related

[5]Our choice of configurations "PROTES-1" and "PROTES-2" corresponds to the parameters used in the first and second versions of the original work [2].

Table 7. The best generated hallucinogens for the DeepL translator for each requested batch. Results for the two optimizer configurations with batch size 50 (PROTES 1) and 100 (PROTES 2) are reported.

Requests	PROTES-1			PROTES-2		
	Text	Translation	Loss	Text	Translation	Loss
50	ощуъиъв	Feelings	-25.08		N/A	
100	бфзскйт	bfzskyt	-20.66	ъщущчны	Synopsis	-29.56
150	бреощее	Breaking	-27.50		N/A	
200	гбьъьиэ	gbjie	-24.08	лзйшеже	better	-34.01
250	бёацсжю	boatsjue	-23.15		N/A	
300	зсзгвлэ	ssgvle	-30.42	едущпяз	Going	-31.05
350	ёренщял	fucking	-19.84		N/A	
400	бфйтйвф	bfjtjvf	-23.08	ждкнжюю	waiting for	-32.49
450	иьллтет	yyllt	-18.23		N/A	
500	пслсждб	pslsjdb	-28.03	лоюоыыф	looyouyf	-23.54
550	рбэхеёе	rbhehehehehe	-22.68		N/A	
600	аэждяэй	aejay	-16.74	псжфйбз	psjfybz	-27.24
650	быншийя	former	-34.84		N/A	
700	сахкьй	Sahkyy	-19.91	ёсычвжь	urchin	-42.89
750	кццьаъг	ktsuag	-19.19		N/A	
800	клчочлй	klcholy	-24.74	бкдммсд	bcdmsd	-26.14
850	ёбсышчн	Fucking	-31.27		N/A	
900	йьръжиь	yrzhi	-21.52	щуэёдьу	squeeze	-32.59
950	ёёщеяйк	urchin	-30.73		N/A	
1000	чотёайь	READ MORE	-35.34	счеочье	account	-32.32

distributions for "PROTES-1" and "PROTES-2". As can be seen from the above results, the second optimizer configuration gives better quality results, but in both cases hallucinogens are generated successfully. Thus, our problem of generating adversarial attacks is successfully solved with the default optimizer parameters. However, convergence curves shown in Figure 9 indicate that if there are more impressive budgets for requests to the translator, further improvement of the results is possible.

§4. Related work

In recent years, large language models have produced significant improvements in various NLP areas, especially in generative tasks. A lot of new concepts were introduced, starting from attention mechanism [1], Transformers [28] to multitask, learning from instructions [31] and human feedback [32]. The latter has become extremely popular in the generative context including machine translation. Consequently, the usage of machine translation tools has become a necessary compound for understanding a foreign language. Unfortunately, like other neural network-based algorithms, these tools are vulnerable to adversarial examples [16]. Starting from text classification [14, 19, 20], vulnerability and robustness received a lot of attention in the NLP community. For MT systems one of the pioneering works was [13], where a character-level approach to generate adversarial examples was proposed. Inheriting from HotFlip [15], they considered settings where only a few symbols in an input query are subject to change, imitating typos.

While white-box optimization may yield stronger adversarial perturbations it implies access to the model's architecture and weights which is impractical in the case of online MT tools. The work [29] considered a white-box universal approach to a targeted attack on conditional text generation. The authors modeled perturbation as an insertion of a trigger, a token sequence of small length, that results in a generated sequence similar to the target set of sentences. While during experiments certain triggers cause a model to produce sensitive racist output, they are generally meaningless and similarly to character-level attacks are easy to detect. Authors of [18, 24] reported high attack transferability making this approach promising for black-box setup, however, the research is limited only to the GPT-2 model for generation task. The above papers use greedy techniques to walk through the searching space during the optimization, on the other hand, attacks on NLP models could be found via projection onto embeddings [29], and for MT task this was discovered in [7, 23, 25]. In [33], it was shown that black-box optimization may yield transferable word-level attack that fools online translation tools, e.g., Baidu and Bing. This work proposed to use the word saliency as the measure of uncertainty. Masking candidates the saliency was estimated via additional BERT model [11] which lead to strong readable and imperceptible adversaries, however, neither human evaluation was performed nor quantities results for online tools were given. In [30], a gradient-based approach to generate phrase-level adversarial examples for neural MT systems was proposed. Similarly to [33], it is proposed to estimate the vulnerable word positions are estimated in an input phrase with the use of gradient information and replace corresponding words by the candidates computed with an auxiliary model.

We also note the recent work [17], where the hallucination problem of MT systems is discussed and the method for detecting and alleviating such hallucinations is presented. The authors identified a set of hallucinations in a large number of translations by various hallucination detection methods (anomalous encoder-decoder attention, simple model uncertainty measures, etc.), and gathered for them human annotations. This allowed them to conduct a comparative analysis of detection methods and to suggest a new approach for detection.

§5. Conclusion

In this work, we propose a simple and effective approach to generate hallucinogens, i.e., nonsensical gibberish in one language that is translatable into another language by online translation tools. We evaluated our method on popular online translation systems: Google, DeepL, and Yandex. We have found that such systems process adversarial examples unpredictably: they not only translate nonsensical input in Russian but also cannot translate seemingly meaningful English phrases. This vulnerability may interfere with understanding a new language and worsen user's experience while using machine translation systems; hence, additional improvements of these tools are required to establish better translation.

Acknowledgments

This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Grant No. 075-15-2020-801). AC would like to thank Lev Chertkov for discovering the possibility of successful adversarial attacks on online translators using the translation result for a set of hallucinogens.

References

1. D. Bahdanau, K. Cho, and Y. Bengio, *Neural machine translation by jointly learning to align and translate*, arXiv preprint arXiv:1409.0473 (2014).
2. A. Batsheva, A. Chertkov, G. Ryzhakov, and I. Oseledets, *PROTES: probabilistic optimization with tensor sampling*, arXiv preprint arXiv:2301.12162 (2023).
3. Y. Berger, *Israel arrests Palestinian because Facebook translated "good morning" to "attack them"*, Ha'aretz **22** (2017).
4. M. Blohm, G. Jagfeld, E. Sood, X. Yu, and N. T. Vu, *Comparing attention-based convolutional and recurrent neural networks: Success and limitations in machine*

reading comprehension, Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018, 2018, pp. 108–118.

5. Y. Chen, H. Gao, G. Cui, F. Qi, L. Huang, Z. Liu, and M. Sun, *Why should adversarial perturbations be imperceptible? rethink the research paradigm in adversarial nlp*, arXiv preprint arXiv:2210.10683 (2022).
6. M. Cheng, J. Yi, P. Chen, H. Zhang, and C. Hsieh, *Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples*, The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, 2020, pp. 3601–3608.
7. M. Cheng, J. Yi, H. Zhang, P.-Y. Chen, and C.-J. Hsieh, *Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples*, Proceedings of the AAAI Conference on Artificial Intelligence **34** (2018).
8. A. Chertkov, G. Ryzhakov, G. Novikov, and I. Oseledets, *Optimization of functions given in the tensor train format*, arXiv preprint arXiv:2209.14808 (submitted to IEEE Computing in Science and Engineering) (2022).
9. A. Cichocki, N. Lee, I. Oseledets, A.-H. Phan, Q. Zhao, and D. Mandic, *Tensor networks for dimensionality reduction and large-scale optimization: Part 1 low-rank tensor decompositions*, Foundations and Trends in Machine Learning **9** (2016), no. 4-5, 249–429.
10. A. Cichocki, A. Phan, Q. Zhao, N. Lee, I. Oseledets, M. Sugiyama, and D. Mandic, *Tensor networks for dimensionality reduction and large-scale optimization: Part 2 applications and future perspectives*, Foundations and Trends in Machine Learning **9** (2017), no. 6, 431–673.
11. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, arXiv preprint arXiv:1810.04805 (2018).
12. J. Ebrahimi, D. Lowd, and D. Dou, *On adversarial examples for character-level neural machine translation*, Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018, 2018, pp. 653–663.
13. J. Ebrahimi, D. Lowd, and D. Dou, *On adversarial examples for character-level neural machine translation*, arXiv preprint arXiv:1806.09030 (2018).
14. J. Ebrahimi, A. Rao, D. Lowd, and D. Dou, *Hotflip: White-box adversarial examples for text classification*, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers, 2018, pp. 31–36.
15. J. Ebrahimi, A. Rao, D. Lowd, and D. Dou, *HotFlip: White-box adversarial examples for text classification*, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (Melbourne, Australia), Association for Computational Linguistics, July 2018, pp. 31–36.
16. I. J. Goodfellow, J. Shlens, and C. Szegedy, *Explaining and harnessing adversarial examples*, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

17. N. M. Guerreiro, E. Voita, and A. F. Martins, *Looking for a needle in a haystack: a comprehensive study of hallucinations in neural machine translation*, arXiv preprint arXiv:2208.05309 (2022).
18. C. Guo, A. Sablayrolles, H. Jégou, and D. Kiela, *Gradient-based adversarial attacks against text transformers*, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (Online and Punta Cana, Dominican Republic), Association for Computational Linguistics, November 2021, pp. 5747–5757.
19. J. Li, S. Ji, T. Du, B. Li, and T. Wang, *Textbugger: Generating adversarial text against real-world applications*, ArXiv **abs/1812.05271** (2018).
20. L. Li, R. Ma, Q. Guo, X. Xue, and X. Qiu, *BERT-ATTACK: Adversarial attack against BERT using BERT*, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Online), Association for Computational Linguistics, November 2020, pp. 6193–6202.
21. I. Oseledets, *Tensor-train decomposition*, SIAM Journal on Scientific Computing **33** (2011), no. 5, 2295–2317.
22. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., *Language models are unsupervised multitask learners*, OpenAI blog **1** (2019), no. 8, 9.
23. S. Sadrizadeh, A. D. Aghdam, L. Dolamic, and P. Frossard, *Targeted adversarial attacks against neural machine translation*, ArXiv **abs/2303.01068** (2023).
24. S. Sadrizadeh, L. Dolamic, and P. Frossard, *Block-sparse adversarial attack to fool transformer-based text classifiers*, ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 7837–7841.
25. ______, *TransFool: an adversarial attack against neural machine translation models*, arXiv preprint arXiv:2302.00944 (2023).
26. K. Sozykin, A. Chertkov, R. Schutski, A.-H. Phan, A. Cichocki, and I. Oseledets, *TTOpt: a maximum volume quantized tensor train-based optimization and its application to reinforcement learning*, Advances in Neural Information Processing Systems, 2022.
27. C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, *Intriguing properties of neural networks*, 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014.
28. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, *Attention is all you need*, Advances in neural information processing systems **30** (2017).
29. E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh, *Universal adversarial triggers for attacking and analyzing NLP*, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (Hong Kong, China), Association for Computational Linguistics, November 2019, pp. 2153–2162.
30. J. Wan, J. Yang, S. Ma, D. Zhang, W. Zhang, Y. Yu, and Z. Li, *Paeg: Phrase-level adversarial example generation for neural machine translation*, Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 5085–5097.
31. Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Naik, A. Ashok, A. S. Dhanasekaran, A. Arunkumar, D. Stap, et al., *Super-natural instructions: Generalization via declarative instructions on 1600+ NLP tasks*, Proceedings of

the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 5085–5109.

32. Z. J. Wang, D. Choi, S. Xu, and D. Yang, *Putting humans in the natural language processing loop: A survey*, arXiv preprint arXiv:2103.04044 (2021).
33. X. Zhang, J. Zhang, Z. Chen, and K. He, *Crafting adversarial examples for neural machine translation*, 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), 2021, pp. 1967–1977.

Поступило 6 сентября 2023 г.

Skolkovo Institute
of Science and Technology,
Moscow, Russia;
Institute of Numerical Mathematics,
Russian Academy of Sciences
E-mail: `a.chertkov@skoltech.ru`

Moscow Institute
of Physics and Technology,
Moscow, Russia;
Sber AI Lab, Moscow, Russia
E-mail: `tsimboy.oa@phystech.edu`

Skolkovo Institute
of Science and Technology,
Moscow, Russia
E-mail: `mikhail.pautov@skoltech.ru`

Skolkovo Institute
of Science and Technology,
Moscow, Russia;
Institute of Numerical Mathematics,
Russian Academy of Sciences;
AIRI, Moscow, Russia
E-mail: `i.oseledets@skoltech.ru`