Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2023"

June 14–16, 2023

Parameter-Efficient Tuning of Transformer Models for Anglicism Detection and Substitution in Russian

Daniil Lukichev
HSE University, Sber
Moscow, Russia
peroprosi@gmail.com

Darya Kryanina
HSE University
Moscow, Russia
daryd388@gmail.com

Anastasia Bystrova
HSE University
Moscow, Russia
eyer89@gmail.com

Alena Fenogenova
SberDevices
Moscow, Russia
alenush93@gmail.ru

Maria Tikhonova
HSE University, SberDevices
Moscow, Russia
m_tikhonova94@mail.ru

Abstract

This article is devoted to the problem of Anglicisms in texts in Russian: the tasks of detection and automatic rewriting of the text with the substitution of Anglicisms by their Russian-language equivalents. Within the framework of the study, we present a parallel corpus of Anglicisms and models that identify Anglicisms in the text and replace them with the Russian equivalent, preserving the stylistics of the original text.

Keywords: Anglicisms, paraphrase, natural language processing, machine learning, language models, style-transfer

DOI: 10.28995/2075-7182-2023-22-295-306

Эффективное по числу обучаемых параметров обучение трансформерных моделей для задач детекции и замены англицизмов в русском языке

Лукичев Даниил
НИУ ВШЭ, Sber
Москва, Россия
peroprosi@gmail.com

Крянина Дарья
НИУ ВШЭ
Москва, Россия
daryd388@gmail.com

Быстрова Анастасия
НИУ ВШЭ
Москва, Россия
eyer89@gmail.com

Феногенова Алена
SberDevices
Москва, Россия
alenush93@gmail.ru

Тихонова Мария
НИУ ВШЭ, SberDevices
Москва, Россия
m_tikhonova94@mail.ru

Аннотация

Данная статья посвящена проблеме англицизмов в текстах на русском языке: задачам детекции и автоматического переписывания текста с заменой англицизмов на их русскоязычные аналоги. В рамках исследования мы представляем параллельный корпус, а также модель, которая выявляет англицизмы в тексте и заменяет их на русский эквивалент, сохраняя стилистику исходного текста.

Ключевые слова: англицизмы, парафраз текста, обработка естественного языка, машинное обучение, языковые модели, стилевой трансфер

1 Introduction

Language reflects the society to which it belongs. Its lexis reflects undergoing changes in political, scientific, technological, and other spheres of life. As new scientific and technical inventions emerge regularly, new words (neologisms) are coined to denote new concepts. The English language has a vast influence in the context of globalisation, exerted by global economic, social, and cultural processes over national ones. “The English language finds itself at the centre of the paradoxes which arise from globalisation. It provides the lingua franca essential to the deepening integration of global service-based economies. It facilitates transnational encounters and allows nations, institutions, and individuals worldwide to communicate their world view and identities“ (Graddol, 2006).

English nowadays is an international language of communication, business, education, and innovation. English has affected most languages in the past 100 years (Görlach, 2002b). For this reason, Görlach (2002a) called the English language “the world’s biggest lexical exporter“ , as most of the newly-coined words are English. Moreover, statistics show that 14.7 English neologisms are created per day[1], making English a highly productive *Source Language* (or SL, in short).

A significant number of English words are integrated into different spheres of human activity (e.g., modern and youth culture, civil and political life, IT, science, education, sports, medicine) in the form of loanwords. English borrowings (or Anglicisms), thus, form a vast lexical stratum in many languages, including Russian. However, often the meaning of these loanwords is uncertain or domain-specific and incomprehensible to people outside a particular field or social strata. Therefore, Anglicisms may impede effective communication between representatives of different generations, professions, subcultures. Furthermore, Anglicisms are inappropriate in some official and scientific discourse unless they “refer to terminology or common vocabulary recorded by explanatory dictionaries of the Russian language“ (Апетян, 2011). In this regard, we frequently have to adjust our writing and speaking styles to a particular audience, social context, or formality of the occasion. In addition, the Anglicisms detection and substitution task is relevant in *Natural Language Processing* (or NLP, in short). Anglicisms often pose challenges for this sphere (for example, machine translation, rewriting and summarization, text-to-speech) as many systems are often dependent on the lexicon.

This paper presents methods for automatic Anglicism detection and their elimination via paraphrasing the original text with these loanwords replaced by their native equivalents. These methods can contribute to many NLP systems enhancing the accuracy of large language models or machine translation systems. Moreover, they can make contribution to language correction and proofreading applications. By identifying potential loanwords, the Anglicism detector can assist writers and editors in to ensure grammatical and stylistic accuracy in written content. Altogether, our models can improve the text’s overall readability by replacing Anglicisms with more natural and understandable phrases in the target language. Such tools can be particularly useful in business, education, science, and journalism, where clear and effective communication is crucial. In addition, we present a parallel corpus of Anglicisms in Russian[2] and the code is available on our GitHub repository[3].

Thus, the contribution of our paper is three-fold: (I) first, we present a parallel corpus for the Anglicisms detection and their substitution with the detailed Anglicism markup, (II) we train and evaluate several models for Anglicisms detection (III) we present, several generation models for Anglicisms substitution.

The rest of the paper is structured as follows: in section 2, we overview the papers related to this research. Next, in section 3, we formally define the task. Section 4 describes the Anglicism dataset, section 5 discusses the methods we used, section 6 describes the metrics we used and the experimental setup, and section 7 presents evaluation results. Finally, section 8 concludes the paper.

2 Related work

The task of Anglicisms detection is relevant in NLP research: these words often refer to out-of-vocabulary words, and as many systems are often dependent on the lexicon, it poses various problems for machine translation, text processing, speech recognition, Natural Language Understanding, and text-to-speech synthesis (Jawahar et al., 2021), (Weller et al., 2022) (Pritzen et al., 2021). And the global trend

[1] `https://languagemonitor.com/`

[2] `https://huggingface.co/datasets/shershen/ru_anglicism`

[3] `https://github.com/dalukichev/anglicism_removing`

is gaining momentum: code-switching (the mixing of languages within a single conversation or text), the predominance of Anglicisms over the *Receptor Language* (or RL, in short) equivalents, the emergence of hybrid languages (e.g., Frenglish, Denglisch, Runglish, or Spanglish).

There are multiple works related to Anglicisms detection in different languages, e.g. detecting Anglicisms in Spanish (Álvarez Mellado and Lignos, 2022). The article describes the creation of an annotated corpus of Spanish text containing examples of unassimilated borrowings, which can be used to train machine learning models to identify such borrowings in new texts. The corpus has 370,000 tokens. The authors also propose several approaches to modelling unassimilated borrowings, including machine learning algorithms such as decision trees, support vector machines and rule-based systems that rely on linguistic features such as phonetics, morphology, and syntax. CRF, BiLSTM-CRF, and Transformer-based models were used to assess their performance on a new annotated corpus of Spanish newswire full of unassimilated lexical borrowings. The results of this work demonstrate that a BiLSTM-CRF model beats results produced by a multilingual BERT-based model.

Another idea for borrowed word detection is presented in (Miller et al., 2020), where the authors focus on phonological and phonotactic aspects of words in a language for the detection in monolingual word lists using such methods as Markov Models, Bag of Sounds and Neural Networks. The authors presented the idea to train a lexical language model on a dataset of annotated borrowings and then use it in detection for previously unseen word loans. The model performed well when tested on artificially generated words, but the three methods proved ineffective on a sample of actual words taken from WOLD [4]. Failure analysis shows that to achieve a positive result in the detection task, many borrowed words from a given language and coherent and consistent word properties are required. For our task, this problem was also taken into account.

Detecting Anglicisms in the Russian language has some peculiar features due to their transliteration into the Cyrillic script (comparison: Youtube [en] - ютьюб/ютуб [ru]; big data [en] - биг дата [ru]), lexicalization and some internal processes in the language (loanwords constitute an effective mechanism for word formation). The authors (Fenogenova et al., 2016) proposed an automated method for Cyrillic-written Anglicism detection based on the idea that speakers tend to preserve phonetic and orthographic properties of the borrowed words. The proposed method involves a combination of two approaches: 1) a linguistic approach based on identifying patterns of English words in Russian text, and 2) a machine learning approach that utilises a feature-based classifier to predict whether a given word is an Anglicism. Using transliteration (ru-en), phonetic transcribing(en-ru) and morphological analysis methods and various filters, authors compose a list of "unknown Anglicism" pairs. They used the Levenshtein distance (Levenshtein and others, 1966) with thresholds (2-3) to measure the similarity between two words in a pair, and the possible candidates' shortlist was created. With the help of Skip-Gram and CBOW, the list of hypotheses was shortened: if words are semantically and phonetically similar and are close in the word2vec model, they can be considered borrowings.

The substitution of Anglicisms in a text can be viewed as a paraphrasing task. In research mentioned in (Egonmwan and Chali, 2019), the authors present a new method for text paraphrasing based on the seq2seq and Transformer-based (Vaswani et al., 2017) models. As a result, the authors proposed a new TRANSEQ framework that combines the efficiency of the transformer model and seq2seq and improves the current state-of-the-art (Gupta et al., 2017) of QUORA and MSCOCO paraphrase data.

In our work, we trained the models for Anglicisms detection and their substitution using different variations of prompt-tuning techniques. The prompt-tuning method was proposed in (Lester et al., 2021). The fundamental concept of this approach involves training soft prompts, which are incorporated into the input sequence passed to the model while all other parameters of the model are frozen.

This idea was further developed in (Liu et al., 2021), where the authors introduce the concept of deep prompt tuning, which involves adding prompts in different layers as prefix tokens. In (Konodyuk and Tikhonova, 2022), the authors studied the applicability of the prompt-tuning method for the Russian language: they showed that it could be a good alternative to model training techniques.

In addition, in our research, we experiment with low-rank adaptation methods (or LoRA) proposed in

[4] World Loanword Database: `https://wold.clld.org`

(Hu et al., 2021). This method compresses the original language model into a low-rank representation that captures the essential information for the target task. This compression is achieved through a low-rank matrix factorization, which decomposes the original weight matrices of the model into two low-rank matrices. Once the low-rank representation of the original language model is obtained, the compressed model is fine-tuned on the target task using a small amount of labelled data. The fine-tuning process updates the compressed parameters of the model to suit the target task better while preserving the most important information from the original model. The authors demonstrated the effectiveness of the LoRA method in several NLP tasks. In addition, they showed that the LoRA approach generates compressed models that exhibit significantly smaller sizes than the original models while still achieving comparable or better performance on the target tasks.

3 Task Definition

In this paper, we formulate the Anglicism substitution (or elimination) problem as the task of rewriting a sentence by replacing Anglicisms with their Russian equivalents.

In our work, we define an Anglicism based on the definition of Görlach(Görlach, 2002b): “a word or idiom that is recognizably English in its form (spelling, pronunciation, morphology, or at least one of the three), but is accepted as an item in the vocabulary of the receptor language”.

According to Pulcini (2012), there are different types of lexical borrowings:

1. **phrasal borrowings**: usually multi-word units or whole phrases, i.e. collocations, idioms, proverbs. (*e.g.,* “она, конечно, *бест оф зе бест*” (best of the best), “*ху из ху*” (who is who)).
2. **lexical borrowings**: words or multi-word units.
   - (a) *direct*: formal evidence of the SL is detectable.
     - i. loanword – borrowed from SL; meaning in RL is close to meaning in SL (*e.g.*, *голкипер* - goalkeeper, *нон-стоп* - non-stop)
     - ii. hybrid – a combination of SL and RL elements (*e.g.,* (OVER-) + adv./adj.: *овер-дофига* домашки, *овер-пресный* рассказ)
   - (b) *indirect*: the SL model is reproduced in the RL through native elements.
     - i. Calques – reproduce the etymon in the form and meaning or meaning only.
       - A. loan translation – translation of SL item into RL (*e.g.,* небоскреб - skyscraper, утечка мозгов - brain drain, промывка мозгов - brainwashing);
       - B. loan rendition – compound or multi-word unit, one part of which is translated from SL and the other is a loose equivalent of the SL part (e.g., *топовый* (TOP + овый: adj.affix) блогер, *оффлайновое* (OFFLINE + овое: adj.affix) издание, *фолловить* (FOLLOW + ить: verb.affix) звезду, *фаниться* (FUN + ить +ся: verb refl.affix));
       - C. loan creation – RL freely renders the SL equivalent (*e.g.*, синий чулок - blue stocking).
     - ii. Semantic loans - an already existing item in the RL takes a new meaning after a SL one. (*e.g.*, обои (на экране) - wallpaper, карта - bank card)

In addition, it is noteworthy to mention such a phenomenon as *Pseudo-Anglicisms*, which are either:

- lexical units borrowed from English into another language, which have a meaning differing from the SL, and which are used in contexts and situations in which they would never appear in English (смокинг(smoking) -> dinner jacket, автостоп (autostop) -> hitch-hiking, паркинг(parking) -> parking lot));
- Russian formations created by combining English morphemes or imitating the phonetic shape of English words (e.g., фейс контроль - “face control”, рекордсмен - “recordsman” - (record holder) (Дьяков, 2012).

In this paper, both Anglicisms and pseudo-Anglicisms are the objects of our interest. Therefore, examples of pseudo-Anglicisms were included in the dataset along with Anglicisms (for simplicity, we refer to both types simply as Anglicisms).

Borrowed words, as was mentioned earlier, are altered to fit the phonetic and grammatical structure of the language. As English and Russian employ different alphabetic systems, loanwords from English are transliterated into the native Cyrillic-based writing system, where Anglicisms usually adopt the structure of the English source word and typically have the set of endings presented in Table 1.

-ер [-er]	спикер [speaker], бартер [bartender], стриммер [streamer]
-инг [-ing]	консалтинг [consulting]
-мен [-man]	спортсмен [sportsman]
-мент [ment]	энтертеймент [entertainment], истеблишмент [establishment]
-ист [-ist]	активист [activist], лоббист [lobbyist]
-зер [-ser]	мерчендайзер [merchandiser], тизер [teaser]
-изм [-ism]	расизм [racism], нарциссизм [narcissism]
-енд(энд) [-end]	уикэнд [weekend], хэппиэнд [happy end], бэкенд [backend]
-аут [-out]	таймаут [time out], камингаут [coming out], чилаут [chill out]
-ент/ант [-ent]	оппонент [opponent], резидент [resident], фигурант [figurant]
-джер [-ger]	мессенджер [messenger], тинейджер [teenager]
-бэк [-back]	флешбэк [flashback], фидбэк [feedback], хэтчбэк [hatchback]

Table 1: Anglicism endings in Russian

In the Russian language, Anglicisms usually undergo a process known as **domestication**, which poses challenges to NLP systems due to the lack of standardization and inconsistency in the usage of domesticated and non-domesticated borrowings. Domestication refers to how a language adapts foreign words or expressions to fit into its linguistic system, making them sound more natural and familiar to native speakers. This process is usually accompanied by altering the word's spelling, pronunciation, or meaning to better fit into the RL's linguistic system. In addition, the borrowed word is altered to fit the phonetic and grammatical structure of the language. For example, *софт* (software); "грозятся закидать *дизами*" (dislikes); "нужно установить обнову на *винду*" (Windows).

4 Data

To create an Anglicisms dataset, we collected 1084 sentences which contained 472 unique words from different domains. This data was collected semi-automatically from several sources (the Russian National Corpus[5], dictionaries (e.g., A.I. Dyakov's[6], dictionary of Anglicisms in Russian language, Russian Wikidictionary[7]), several Internet resources such as Kartaslov[8], Habr[9], Pikabu[10], as well as blogs and social media sources.

To create a parallel corpus, we paraphrased each sentence replacing all Anglicisms with their Russian equivalents, which were taken from multilingual dictionaries[11],[12] and Wikipedia[13]. All sentences were validated and paraphrased manually by the linguists. It should also be noted that replacing an Anglicism with a single word was not always possible. In some cases, they were substituted with collocations or set expressions (фидбэк (feedback) - обратная связь, краудфандинг (crowdfunding) - коллективный сбор средств, фандрайзинг (fundraising) - сбор средств, оффер (job offer) - предложение по трудоустройству, приглашение на работу).

Thus, we obtained a novel corpus for Anglicisms detection and substitution in the Russian Language[14]. It consists of parallel text pairs: an original sentence with Anglicisms and a sentence in which their Rus-

[5] https://ruscorpora.ru

[6] http://Anglicismdictionary.ru

[7] https://ru.wiktionary.org/wiki/РӘРөСЂРҳРүР«СЃРҷCS:Р№РҳР«РњР«РүРҷРчРёСЌ/ru

[8] https://kartaslov.ru

[9] https://habr.com/

[10] https://pikabu.ru/

[11] Multitran: https://www.multitran.com/

[12] Cambridge dictionary: https://dictionary.cambridge.org/dictionary/english-russian/

[13] https://ru.wikipedia.org/wiki/

[14] https://huggingface.co/datasets/shershen/ru_anglicism

Word	**Form**	**Sentence**	**Paraphrase without Anglicisms**
агриться	сагрилась	Пойдем пока она не сагрилась на нас.	Пойдем пока она н не разозлилась на нас.
кринж	кринжового	Ничего более кринжового я в жизни не видел.	Ничего более постыдного я в жизни не видел.
трушный	трушным	Рядом с тобой даже Джонни Бой был трушным пацаном.	Рядом с тобой даже Джонни Бой был настоящим пацаном.
слот, позер	слоты, позеры	Во дворе эти позеры заняли все парковочные слоты.	Во дворе эти притворщики заняли все парковочные места.
эпикфейл	эпикфейла	Моему злорадству по поводу эпикфейла сего сайта нет предела.	Моему злорадству по поводу провала сего сайта нет предела.

Table 2: A snippet from the Anglicism dataset.

Sentence (English)
Let's go before she gets angry at us.
That's the most cringe-worthy thing I've ever seen in my life.
Next to you, even Johnny Boy was a real kid.
In the yard, these posers took up all the parking slots.
My gloating over the epic fail of this site has no limits.

Table 3: Anglicism dataset format. Translation of the sentences from Table 2. Due to the Anglicism specifics, both sentences (with and without Anglicisms) are translated into English the same way.

sian analogues replace them. A snippet from the dataset is presented in Table 2 (the English translation of the sentences is given in Table 3).

The resulting dataset consists of 1084 sentence pairs divided into train and test parts (999 for the train part and 85 for the test part). The test part includes 30 unique Anglicisms which are not encountered in the train part.

The modest size of the dataset can be partially explained by the fact that in our work, we decided to prioritize the data quality before its quantity. That coincides with the results of the recent research (Zhou et al., 2023), which shows that a relatively small amount of high-quality data can be more beneficial than large low-quality datasets. Thus, we put additional effort into collecting data and selecting good Anglicism examples, which took additional time and resources. Namely, to ensure the annotation quality and to avoid potential errors, we avoided using such annotation services as Yandex.Toloka[15] and paraphrased all sentences with the help of professional linguists, which was more expensive and time-consuming. As a result, we obtained a relatively modest but high-quality dataset. In addition, it should be noted that we took into account the current dataset size and selected suitable methods, such as prompt-tuning and LoRA (see section 5), which can be successfully applied to such amounts of data (Konodyuk and Tikhonova, 2022).

5 Method

Our approach consists of two parts: 1) a model for Anglicisms detection and 2) a paraphrasing model, which rewrites a sentence, replacing the Anglicisms with their Russian-language equivalents.

5.1 Prompt-tuning

Both parts of the algorithm use different variations of prompt-tuning(Lester et al., 2021). Prompting is a technique that provides additional information to the language model to condition during the generation of output Y. Typically, this is achieved by adding a series of tokens P to the input X, resulting in a new input $[P; X]$. The model's parameters remain fixed while it maximizes the possibility of generating the correct Y:

[15] `https://toloka.yandex.ru`

$$Y = \arg\max Prob_{\theta}(Y|[P; X]).$$

The generative model incorporates the prompt tokens P, into the model's embedding table, parameterized by frozen θ. Finding an optimal prompt involves selecting prompt tokens from a fixed vocabulary of embeddings, either through manual search or non-differentiable search methods. Prompt tuning, on the other hand, enables the prompt to have its own dedicated parameters, θ_p, that can be updated. Prompt tuning involves using a fixed prompt of special tokens, with only the embeddings of these prompt tokens being updatable. In essence, prompt tuning eliminates the requirement for the prompt P to be parameterized by θ, as in traditional prompting.

There are different types of initialization of added embeddings:

1. embeddings of random words from the dictionary
2. embeddings of class labels from the task
3. random initialization (does not work well)

We use this variant of prompt-tuning for the Anglicism substitution part, applied in combination with the paraphrase decoder-based models. In our approach, embeddings of random tokens from the first layer of the model are used.

As for Anglicisms detection, we, among other approaches, deployed advanced prompt-tuning. However, in the original prompt-tuning, only continuous prompts are incorporated into the input embedding sequence, which presents two major drawbacks. First, the sequence length limitations impose constraints on the number of trainable parameters. Secondly, the impact of the input embeddings on model predictions is relatively indirect. To overcome these obstacles, *P-Tuning v2* (Liu et al., 2021) introduces the concept of deep prompt tuning, which involves adding prompts in different layers as prefix tokens. This approach enables tuning more task-specific parameters (between 0.1 and 3 per cent), providing greater per-task capacity while remaining parameter-efficient. Additionally, prompts added to deeper layers have a more direct impact on the model's predictions.

5.2 Anglicism detection

We regard the Anglicism detection problem as a token classification task. Tokens that are Anglicisms are labelled as **1**, and the remaining are labelled as **0**. For this task, we evaluated three models:

- **ruBert-tiny**[16]: a small BERT-like model;
- **ruRoberta-large**[17]: a large Russian language RoBERTa model;
- **XLM-RoBERTa**[18]: a large multilingual RoBERTa model.

Since large models tend to overfit on a small amount of data, we used different approaches for training small and large models. Namely, for small models, we used a relatively low learning rate (see section 6.2 for the details). For the large models, we implemented the P-Tuning v2 technique. In addition, we have incorporated the trained tensors into each model layer, effectively decreasing the number of trainable parameters to prevent overfitting. All together, this enables to fine-tune large models for the downstream task, even with limited data.

5.3 Anglicism substitution

We used the prompt-tuning technique to train a paraphrasing model for Anglicism substitution. The important aspect of this approach is to specify the position of trained embeddings within the model's input. In our work, we used the following types of prompts formats:

- **only sent**: <prompt> sentence with Anglicisms <prompt> its paraphrase without Anglicisms
- **sent + angl**: <prompt> sentence with Anglicisms <prompt> Anglicism <prompt> its paraphrase without Anglicisms

In the first format, the embeddings that have been trained are positioned both at the beginning of the sample and between the sentence and its paraphrase, which does not contain Anglicisms. In the

[16] `https://huggingface.co/cointegrated/rubert-tiny`

[17] `https://huggingface.co/sberbank-ai/ruRoberta-large`

[18] `https://huggingface.co/xlm-roberta-base`

second prompt format, we also pass an Anglicism as a model input together with the original sentence and the sentence paraphrase. For this approach, we need the knowledge of Anglicisms to format our examples. We used an Anglicism detector trained at the Anglicism detection stage. Namely, we utilised ruRoBERTa-large detector, which showed the best results in our experiments on Anglicism detection (see section 6 for the details). Thus, the second approach incorporates two models. The detection model identifies Anglicisms in the sentence and then feeds them, along with trained embeddings, to the input of the paraphrasing model.

We utilise the large-scale Russian language model ruGPT3-Large[19] and a multilingual GPT-based model mGPT[20].

For the low-rank adaptation approach, we add the product of two matrices with dimensions $H \times K$ to all attention layers, where H is the dimension of the hidden state of the model, and K is a small value. In our experiments, we use $K = 4$, which was motivated by the research conducted in (Hu et al., 2021).

6 Experiments

6.1 Evaluation

Anglicism detection As long as we consider the Anglicism detection task as a binary token classification problem, we use binary classification metrics (F1, precision, and recall) for evaluation.

Anglicism substitution As for the Anglicism substitution, we evaluate this part using the following metrics, which are commonly used for generative tasks and the paraphrase tasks in particular:

1. CHRF++[21](Popović, 2015)
2. BLEU score(Papineni et al., 2002)
3. Rouge-L(Lin, 2004)
4. BERTScore(Zhang et al., 2019)
5. LaBSE(Feng et al., 2020)[22]

All metrics listed above are computed between gold paraphrases and model predictions and averaged over the test set.

6.2 Experimental setup

One of the essential hyperparameters of prompt tuning is the length of the prompt. In our research, we use the following prompt lengths:

- *detection*: in our methodology, we introduce prompts of length 100 to each attention layer and optimize them using the learning rate $1e-3$. Additionally, the linear head is optimized with a learning rate of $1e-5$, with a batch size of 8 and for a duration of 10 epochs.
- *sentence-paraphrase approach*: we add a prompt of length 50 before the sentence and a prompt of length 40 between the sentence and the paraphrase. We optimize prompts with a learning rate of $1e-3$ and linear head with a learning rate of $1e-5$ with a batch size of 8 and for 5 epochs.
- *sentence-anglicism-paraphrase approach*: we add a prompt of length 50 before the sentence, a prompt of length 20 between the sentence and the Anglicism and a prompt of length 40 between the Anglicism and the paraphrase. We optimize prompts with a learning rate of 1e-3 and linear head with a learning rate of $1e-5$ with a batch size of 8 and for 5 epochs.

In low-rank adaptation approaches, the models are trained with the learning rate $1e-5$, which is kept the same for both the model and linear head parameters, using a batch size of 8 and for a total of 15 epochs. The AdamW optimizer (Loshchilov and Hutter, 2017) and linear scheduler with warm-up are employed in all the experiments.

[19] `https://huggingface.co/sberbank-ai/rugpt3large_based_on_gpt2`

[20] `https://huggingface.co/sberbank-ai/mGPT`

[21] `https://huggingface.co/spaces/evaluate-metric/chrf`

[22] `https://huggingface.co/sentence-transformers/LaBSE`

7 Results

7.1 Anglicism detection

Analyzing the results of Anglicism detection (see Table 4), it can be observed that ruRoberta-large shows the best quality surpassing other models in all metrics. XLM-RoBERTa also produces competitive results, while ruBert tiny performs much worse. We hypothesize that such low performance can be explained by the fact that the model was fine-tuned without prompt tuning, and even though it contains a small number of parameters, it still began to overfit too quickly on the small dataset.

The obtained results coincide with the work of (Leidig et al., 2014), where the authors tried the combination of several features (G2P confidence, grapheme perplexity, Google hits count) to detect Anglicisms in German and achieved a 0.75 F1 score. The work (Mellado et al., 2021) devoted to the same task for Spanish, presented in IberLef 2021, reported F1 scores ranging from 0.37 to 0.85. In addition, another research for the Norwegian language (Andersen, 2005) is devoted to Anglicism extraction using a combination of methods (rule-based, lexicon-based, and chargram-based). In their work, such a combined approach yielded the most favourable outcome, achieving an overall 0.96 accuracy score for correctly annotated forms and a precision rate of 0.76, which is comparable with our results.

Model	F1	Precision	Recall
ruBERT-tiny (fine-tuning)	0.62	0.59	0.66
ruRoBERTa-large (prompt-tuning)	**0.72**	**0.69**	**0.80**
XLM-RoBERTa (prompt-tuning)	0.70	0.67	0.78

Table 4: Anglicism detection results. Detailed metrics descriptions are given in subsection 6.1.

Besides the general Anglicism detection evaluation, we also performed an additional study of Anglicism detection mistakes. For this, we analyzed the predictions of the best model, that is, the ruRoBERTa-large (prompt-tuning) model (see Table 5 for the most typical mistakes).

Sentence	Model prediction (token level)
В ЛДЦ “Кутузовский” в Москве вы можете пройти полное чек-ап обследование всего организма.	чек-
Если не знаешь как начать дейтиться, то этот коуч научит тебя.	дейт, коуч
Можешь рассчитывать даже на апельсиновый фреш в моём исполнении!	ап

Table 5: Typical Anglicism detection mistakes of the ruRoBERTa-large (prompt-tuning) model.

From the mistake analysis, several conclusions can be made:

1. The model demonstrates a restricted capability in accurately identifying Anglicisms that consist of multiple words connected by hyphens. Although the model can identify such Anglicisms, lowering the sensitivity threshold of the linear classification layer resolves this issue.
2. In the process of tokenization, some Anglicisms are tokenized as several tokens. As a result, the model sometimes marks only the English root as an Anglicism, omitting suffixes and inflections.
3. The model occasionally generates false positive errors by incorrectly marking tokens resembling English word parts as Anglicisms.

7.2 Anglicism substitution

As for the Anglicism substitution results (see Table 6), the two model variants can be highlighted here. Namely, ruGPT3 sent+angl outperforms other models by CHFR++ and BLEU, and ruGPT3 LoRA yields the best score by Rouge-L, BERTscore, and LaBSE. This result was obtained due to the fact that in the first approach, the model did not always replace Anglicism in the sentence. In contrast, in the second approach, the model replaced Anglicism more often, but sometimes not with the same word as in our golden paraphrase. Nevertheless, the substitution the model proposed was semantically close to the golden one. Therefore, metrics measuring semantic proximity, BERTScore and LaBSE turned out to be higher in the second approach. The low-rank adaptation approach has demonstrated its efficiency as it maximizes the potential of large pre-trained models by optimizing all model layers, albeit in a specific manner. The hypothesis that multilingual models cope better with Anglicisms detection and substitution has not been confirmed.

It should also be noted that we solve the Anglicism substitution problem as the generative task and, therefore, employ generative metrics for their evaluation. Thus, due to the possible plurality of the correct answers and the variety of generated output and distinctiveness, these metrics are not expected to reach the theoretical maximum when assessing the effectiveness of generative models like the one in our approach.

Model	**CHRF++**	**BLEU**	**Rouge-L**	**BERTScore**	**LaBSE**
ruGPT3 only sent	0.79	0.58	0.74	0.89	0.91
ruGPT3 sent+angl	**0.81**	**0.72**	0.77	0.91	0.93
mGPT3 only sent	0.75	0.64	0.73	0.89	0.92
mGPT3 sent+angl	0.78	0.68	0.75	0.90	0.91
ruGPT3 LoRA	0.76	0.67	**0.8**	**0.92**	**0.94**
mGPT3 LoRA	0.71	0.62	0.78	0.90	0.91

Table 6: Anglicism substitution results. Detailed metrics descriptions are given in subsection 6.1.

Analyzing the predictions of ruGPT3 Lora, which yielded the best scores by most of the metrics, two main types of mistakes can be highlighted:

1. The model leaves the sentence unchanged. This usually happens with uncommon Anglicisms, which are, by being rare, tokenized into several tokens. For example, in the sentence "Футболист Лионель Месси является амбассадором Adidas." the Anglicism "амбассадором" is tokenized into four tokens, and the model fails to replace it.
2. The model replaces an Anglicism with a wrong word changing the meaning (e.g., "Она скринит наши переписки." paraphrased as "Она проверяет наши переписки."). This is most likely due to the fact that the model failed to learn the correct meaning of the Anglicism.

8 Conclusion

This article is devoted to Anglicism detection in Russian and their substitution with Russian equivalents to ensure effective communication across various social and professional strata. In this work, we presented a parallel corpus of Anglicism, several models for Anglicism detection and a set of generative models for Anglicism substitution. In addition, we compared a series of experiments and performed a comprehensive model evaluation. All the code and all the models are available in our repository[23] and the dataset can be downloaded[24] from HuggingFace project.

As a part of future work, we plan to augment the existing dataset with both new Anglicisms and new sentences with the current one. We hope that such data augmentation will improve the result.

8.1 Possible Misuse

We believe that our research should not be involved in creating content that affects the individual or communal well-being in any way, including

- legislative application or censorship;
- mis- and disinformation;
- infringement of the rights of access to information.

[23] https://github.com/dalukichev/anglicism_removing

[24] https://huggingface.co/datasets/shershen/ru_anglicism

8.2 Biases and data quality

The Anglicism corpus includes large segments representing the Internet domain, and therefore, it may possibly contain a variety of stereotypes and biases. Proper evaluation is still needed to explore possible model vulnerabilities in terms of generalizing on the new data and specific new data.

References

Gisle Andersen. 2005. Assessing algorithms for automatic extraction of anglicisms in norwegian texts. 01.

Elozino Egonmwan and Yllias Chali. 2019. Transformer and seq2seq model for paraphrase generation, November.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding.

Alena Fenogenova, Ilia Karpov, and Viktor Kazorin. 2016. A general method applicable to the search for anglicisms in russian social network texts. // *2016 IEEE Artificial Intelligence and Natural Language Conference (AINL)*, P 1–6. IEEE.

Manfred Görlach. 2002a. *An annotated bibliography of European anglicisms*. OUP Oxford.

Manfred Görlach. 2002b. *English in Europe*. OUP Oxford.

David Graddol. 2006. *English next*, volume 62. British council London.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2017. A deep generative framework for paraphrase generation.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. lora. *CoRR*, abs/2106.09685.

Ganesh Jawahar, El Moatez Billah Nagoudi, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2021. Exploring text-to-text transformers for English to Hinglish machine translation with synthetic code-mixing. // *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, P 36–46, Online, June. Association for Computational Linguistics.

Nikita Konodyuk and Maria Tikhonova. 2022. Continuous prompt tuning for russian: how to learn prompts efficiently with rugpt3? // *Recent Trends in Analysis of Images, Social Networks and Texts: 10th International Conference, AIST 2021, Tbilisi, Georgia, December 16–18, 2021, Revised Selected Papers*, P 30–40. Springer.

Sebastian Leidig, Tim Schlippe, and Tanja Schultz. 2014. Automatic detection of anglicisms for the pronunciation dictionary generation: A case study on our german it corpus. 05.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. // *Soviet physics doklady*, volume 10, P 707–710. Soviet Union.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. // *Text summarization branches out*, P 74–81.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *CoRR*, abs/2110.07602.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam.

Elena Mellado, Luis Espinosa-Anke, Julio Arroyo, Constantine Lignos, and Jordi Porta. 2021. Overview of adobo 2021: Automatic detection of unassimilated borrowings in the spanish press, 10.

John E. Miller, Tiago Tresoldi, Roberto Zariquiey, César A. Beltrán Castañón, Natalia Morozova, and Johann-Mattis List. 2020. Using lexical language models to detect borrowings in monolingual wordlists, Dec.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. // *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, P 311–318.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. // *Proceedings of the Tenth Workshop on Statistical Machine Translation*, P 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Julia Pritzen, Michael Gref, Christoph Schmidt, and Dietlind Zühlke. 2021. A comparative pronunciation mapping approach using g2p conversion for anglicisms in german speech recognition. // *Speech Communication; 14th ITG Conference*, P 1–5. VDE.

Virginia Pulcini, Cristiano Furiassi, and Félix Rodríguez González. 2012. The lexical influence of english on european languages. *The anglicization of European lexis*, 1.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Orion Weller, Matthias Sperber, Telmo Pires, Hendra Setiawan, Christian Gollan, Dominic Telaar, and Matthias Paulik. 2022. End-to-end speech translation for code switched speech. // *Findings of the Association for Computational Linguistics: ACL 2022*, P 1435–1448, Dublin, Ireland, May. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

Elena Álvarez Mellado and Constantine Lignos. 2022. Detecting unassimilated borrowings in spanish: An annotated corpus and approaches to modeling.

Светлана Геннадьевна Апетян. 2011. Англицизмы в структуре масс-медийного и официально-делового дискурсов (лексико-семантический и когнитивно-прагматический аспекты).

Анатолий Иванович Дьяков. 2012. УРОВНИ ЗАИМСТВОВАНИЯ АНГЛИЦИЗМОВ В РУССКОМ ЯЗЫКЕ. *Известия Южного федерального университета. Филологические науки*, (2):113–124.