Natural Language Processing (2024), 1–26
doi:10.1017/nlp.2024.38

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

CoAT: Corpus of artificial texts

Tatiana Shamardina[1,*], Marat Saidov[2,*], Alena Fenogenova[2,*], Aleksandr Tumanov[2], Alina Zemlyakova[2], Anna Lebedeva[2], Ekaterina Gryaznova[2], Tatiana Shavrina[3], Vladislav Mikhailov[4,†,‡] and Ekaterina Artemova[5,†,‡]

[1]ABBYY, Milpitas, CA, USA, [2]HSE University, Moscow, Russia, [3]Institute of Linguistics, RAS, Moscow, Russia, [4]University of Oslo, Oslo, Norway, and [5]Toloka AI
Corresponding author: Vladislav Mikhailov; Email: vladism@ifi.uio.no

(Received 1 June 2023; revised 19 May 2024; accepted 19 May 2024)

Abstract

With recent advances in natural language generation, risks associated with the rapid proliferation and misuse of generative language models for malicious purposes steadily increase. Artificial text detection (ATD) has emerged to develop resources and computational methods to mitigate these risks, such as generating fake news and scientific article reviews. This paper introduces corpus of artificial texts (CoAT), a large-scale corpus of human-written and generated texts for the Russian language. CoAT spans six domains and comprises outputs from 13 text generation models (TGMs), which differ in the number of parameters, architectural choices, pre-training objectives, and downstream applications. We detail the data collection methodology, conduct a linguistic analysis of the corpus, and present a detailed analysis of the ATD experiments with widely used artificial text detectors. The results demonstrate that the detectors perform well on the seen TGMs, but fail to generalise to unseen TGMs and domains. We also find it challenging to identify the author of the given text, and human annotators significantly underperform the detectors. We release CoAT, the codebase, two ATD leaderboards, and other materials used in the paper.

Keywords: text classification; artificial text detection; natural language generation evaluation

1. Introduction

Disclaimer: Parts of this section highlighted in italics are generated by ChatGPT to illustrate the need for facilitating research in the detection of neural texts. We guarantee that the generated text contains no misinformation and provide it solely for illustration purposes.

Modern text generation models (TGMs) excel at producing text that can be indistinguishable from human-written texts, judging by its fluency, coherence, and grammar (Zellers *et al.* 2019; Radford *et al.* 2019). While advanced TGMs are useful for many real-world applications, such as text summarisation or machine translation, their risks are viewed as critical (Weidinger *et al.* 2021; Bommasani *et al.* 2021). Here are some of the most significant ones:

1. *Biases and Stereotypes: Large-scale TGMs are trained on vast amounts of data, which means that they can replicate existing biases and stereotypes that are present in the data. For example, if the*

*These authors are contributed equally to this work.
†Equal advising.
‡Work done while at HSE University.

training data contain gender bias or racial bias, the text generated by the model may also contain such biases.

2. *Misinformation and Fake News: Since TGMs can create convincing and coherent sentences, they can also be used to generate false information and spread misinformation or fake news. This could have serious consequences, such as spreading rumours, influencing elections, and inciting violence*.

3. *Malicious Use: Large-scale TGMs could be used for malicious purposes, such as generating convincing phishing emails, creating fake reviews to manipulate consumer opinion, or even creating convincing fake personas to spread propaganda*.

4. *Ethical Considerations: The use of large-scale TGMs raises ethical considerations around the use of data, privacy, and consent. There are also concerns around the impact of these models on the job market, as they can automate tasks previously performed by humans*.

The text highlighted in italics above is generated by ChatGPT[a] prompted to review risks associated with the rapid development of TGMs in an academic style. This text reads fluently and naturally, adhering to the academic writing norms to a certain degree. It contains non-trivial ideas such as referencing potential impacts on the job market and creating fake personas. Overall, this demonstrates how generated text can be smoothly integrated into an academic article, compromising its authenticity.

The field of *artificial text detection (ATD)* (Jawahar, Abdul-Mageed, and Lakshmanan 2020) aims to develop resources and computational methods to mitigate the risks of misusing TGMs. With advancements of TGMs, the problem has received special interest in the community since humans struggle to distinguish between natural and neural texts (Gehrmann, Strobelt, and Rush 2019; Ippolito *et al.* 2020; Karpinska, Akoury, and Iyyer 2021; Uchendu *et al.* 2021). Detection of artificial texts has been framed in multiple ways, featuring various task formulations and labelling schemes. The most standardised task is a binary classification problem with the goal of determining if the text is automatically generated or not (Adelani *et al.* 2020; Bahri *et al.* 2021). Uchendu *et al.* (2020) studied the neural authorship attribution task aimed to single out one TGM that generated the text. Dugan *et al.* (2023) formulate the boundary-detection task: detect a change point in the text, where a natural text transitions into a neural one.

Although ATD is rapidly developing, there is still a need for creating resources for non-English languages that account for the diversity of the TGMs, natural language generation tasks, and text domains (Uchendu *et al.* 2022). In this paper, we introduce corpus of artificial texts (CoAT), a large-scale ATD corpus for Russian composed of human-written texts from publicly available resources and artificial texts produced by 13 TGMs, varying in the number of parameters, architecture choices, pre-training objectives, and downstream applications. Each TGM is fine-tuned for one or more of six natural language generation tasks, ranging from paraphrase generation to text summarisation. CoAT provides two task formulations and public leaderboards: (i) detection of artificial texts, i.e., classifying if a given text is machine-generated or human-written; (ii) authorship attribution, i.e., classifying the author of a given text among 14 candidates. The design of our corpus enables various experiment settings, ranging from analysing the dependence of the detector performance on the natural language generation task to the robustness of detectors towards unseen TGMs and text domains.

Contributions. Our main contributions are the following: (i) We create CoAT, a large-scale corpus for artificial text detection in Russian (Section 3). (ii) We present a linguistic analysis of the corpus, focusing on the distribution of stylometric features in human-written and machine-generated texts (Section 4). (iii) We provide a detailed analysis of human annotators, non-neural,

[a]Access date: May 8, 2023.

https://doi.org/10.1017/nlp.2024.38 Published online by Cambridge University Press

and transformer-based artificial text detectors in five experiment settings (Section 5). (iv) We release CoAT,[b] source code,[c] and human evaluation project and provide two public leaderboards on the Kaggle platform[d,e].

2. Related work

2.1 Datasets and benchmarks

The community has put much effort into creating datasets and benchmarks for ATD tasks that cover various domains and architectures of TGMs. The design generally includes collecting human-written texts from publicly available sources and generating synthetic texts with a specific decoding strategy by (i) prompting a pretrained TGM without domain adaptation (e.g., with a news article title; Uchendu *et al.* 2020, 2021), (ii) prompting a fine-tuned TGM (i.e., after continuing the pre-training on texts from the target domain; Gupta *et al.* 2020), and (iii) using a TGM trained or fine-tuned for a particular text generation task (e.g., MT; Aharoni, Koppel, and Goldberg 2014).

The GPT-2 output dataset is one of the first to address the detection of texts produced by modern large-scale TGMs (Radford *et al.* 2018). The dataset[f] consists of natural texts from WebText (Reddit) and texts produced by multiple versions of the GPT-2 model fine-tuned on WebText. Munir *et al.* (2021) and Diwan *et al.* (2021) extracted generated text from the subreddit r/SubSimulatorGPT2. Users of this subreddit are GPT-2-based TGMs fine-tuned on the posts and comments from a specific subreddit. The TweepFake dataset (Fagni *et al.* 2021) contains tweets posted by 40 accounts, including statistical and neural TGMs and human users. Adelani *et al.* (2020) propose a dataset for mitigating the generation of fake product reviews using out-of-the-box TGMs and TGMs continuously pretrained on the target domain. Kashnitsky *et al.* (2022), Rodriguez *et al.* (2022), and Liyanage *et al.* (2022) design datasets to detect artificially generated academic content and explore the robustness of trainable detectors towards unseen research domains. The risk of spreading neural fake news and misinformation has facilitated the creation of ATD resources in the news domain, such as the GROVER dataset (Zellers *et al.* 2019), the NeuralNews dataset (Tan, Plummer, and Saenko 2020), and the "All the News" dataset (Gupta *et al.* 2020).

A few recent works explore multi-domain ATD. Bakhtin *et al.* (2019) collect and generate texts from news, multi-genre fictional books, and Wikipedia. Stiff and Johansson (2022) propose a dataset comprising texts from news articles, product reviews, forum posts, and tweets.

Table 1 provides an overview of existing works that have explored artificial text detection for languages other than English. Independent efforts have been made to develop datasets for Bulgarian (Temnikova *et al.* 2023), Chinese (Chen *et al.* 2022), and Spanish (Sarvazyan *et al.* 2023). Wang *et al.* (2024a, 2024c) collect M4, a large-scale multilingual dataset, to test the generalisation abilities of artificial text detectors. Our work differs from related studies in that we (i) create one of the first multi-domain and multi-generator large-scale corpora of artificial texts for Russian, covering standard downstream natural language generation tasks and TGMs; (ii) present a follow-up work on a shared task for artificial text detection in Russian (Shamardina *et al.* 2022), increasing the corpus size and extending the experimental setup to analyse the linguistic properties of human-written and machine-generated texts and explore the robustness of various detectors towards the text domain and TGM size. The earlier CoAT version, namely, the

[b] hf.co/datasets/RussianNLP/coat
[c] github.com/RussianNLP/CoAT
[d] kaggle.com/competitions/coat-authorship-attribution
[e] kaggle.com/competitions/coat-artificial-text-detection
[f] github.com/openai/gpt-2-output-dataset

https://doi.org/10.1017/nlp.2024.38 Published online by Cambridge University Press

Table 1. Artificial text detection datasets and benchmarks for non-English languages. **Notations: H = human-written texts. M = machine-generated texts**

Dataset/ Benchmark	Paper	Language	# texts		# models	# domains
			H	M		
Deepfake-BG	Temnikova *et al.* (2023)	Bulgarian	4,912	4,912	5	1
Generated essays	Chen *et al.* (2022)	Chinese	20,000	20,000	1	1
AuTexTification 1	Sarvazyan *et al.* (2023)	Spanish	24,707	27,484	6	5
RuATD	Shamardina *et al.* (2022)	Russian	107,500	107,500	13	6
M4	Wang *et al.* (2024c)	Chinese	3,000	6,000	2	1
		Urdu	3,000	6,000	2	1
		Arabic	3,000	6,000	2	1
		Indonesian	3,000	6,000	2	1
		Russian	3,000	6,000	2	1
CoAT	(ours)	Russian	123,241	123,241	13	6

RuATD subcorpus, has been included in M4 and used at SemEval-2024 Task 8 on multi-domain, multimodel, and multilingual ATD (Wang *et al.* 2024b, 2024c).

2.2 Artificial text detectors

Feature-based detectors. Classical machine learning methods are widely employed for detecting generated texts. Linear classifiers over TF-IDF character, sub-word, and word N-grams can serve as lightweight and strong baseline detectors. (Manjavacas *et al.* 2017; Solaiman *et al.* 2019; Ippolito *et al.* 2020).Badaskar *et al.* (2008) train detectors over morphological, syntactic, and discourse features, such as POS tags, syntax-based LM's grammaticality scores, and coherence. Another group of features is based on stylometry, a branch of computational linguistics that relies on statistical methods for authorship attribution and analysis of literary style (Holmes, 1994, Abbasi and Chen, 2008; Abbasi and Chen, 2008). Stylometric features are used to train detectors and characterise properties of machine-generated and human-written texts (Uchendu *et al.* 2020, 2021). Specific types of stylometric features can capture issues related to TGMs (Fröhling & Zubiaga, 2021): (i) *lack of syntactic and lexical diversity* (POS tags, named entities, and coreference chains), (ii) *repetitiveness* (N-gram overlap of words and POS tags and counters of stopwords and unique words), (iii) *lack of coherence* (the appearance and grammatical roles of named entities), and (iv) *lack of purpose* (lexical features that represent spatial properties, sentiment, opinion, and logic).

The feature-based detectors are interpretable, cost-effective, and helpful when the dataset size is small (Uchendu, Le and Lee 2023). Stylometric detectors show usefulness in recognising texts generated with certain decoding strategies (Fröhling & Zubiaga, 2021) but are significantly inferior in performance compared to Transformer-based detectors (Schuster *et al.* 2020; Diwan *et al.* 2021; Jones Nurse and Li 2022).

Transformer-based detectors. The current state-of-the-art approach is fine-tuning a pretrained Transformer LM for the ATD classification task. Zellers *et al.* (2019) train a linear layer over the hidden representations from the GROVER and GPT-2 TGMs. RoBERTa (Liu *et al.* 2019) has demonstrated an outstanding performance with respect to many TGMs' configurations and

https://doi.org/10.1017/nlp.2024.38 Published online by Cambridge University Press

domains (Adelani *et al.* 2020; Fagni *et al.* 2021). The TGM-based detectors identify texts generated by the previous TGMs' versions better than by the more recent ones, which confirms that newer TGMs generate more human-like texts. Kushnareva *et al.* (2021) introduce the topological data analysis (TDA) method for ATD, which combines properties of the feature-based and Transformer-based detectors. The TDA features are extracted from the BERT attention matrices that are represented as weighted attention graphs and used to train a linear detector. The features include standard graph properties, descriptive characteristics of barcodes, and distances to attention patterns (Clark *et al.* 2019). The TDA-based detectors outperform TF-IDF-based and neural detectors and show better robustness toward the size of unseen GPT-style TGMs. The Transformer-based detectors are highly effective in ATD tasks and generalisable to out-of-domain sets but computationally expensive.

Zero-shot detectors. An alternative set of methods involves using probability-based measures along with predetermined thresholds (Solaiman *et al.* 2019; Mitchell *et al.* 2023). These methods enable the human-in-the-loop approach, in which a user can recognise whether a text is machine-generated with the assistance of pretrained LMs. The GLTR tool (Gehrmann *et al.* 2019) facilitates interaction between humans and models by presenting statistical characteristics of a text inferred by the model, thereby enhancing the ability of humans to identify artificial texts. MAUVE (Pillutla *et al.* 2021) is a statistical detector that determines the difference in distribution between human and neural texts by utilising the KL-divergence. When MAUVE highlights differences between human and neural texts, human identification of texts generated by GROVER and GPT-2 strongly improves. Dugan *et al.* (2020) propose RoFT (Real or Fake Text), a tool that assists in distinguishing between human-written and machine-generated texts. Their work highlights that TGMs have the ability to deceive humans with just a few sentences. Gallé et al. (2021) explore unsupervised ATD, employing repeated higher-order N-grams. Their findings indicate that certain well-formed phrases occur more frequently in machine-generated texts compared to those written by humans. Although zero-shot detectors tend to underperform compared to the other established methods, they are beneficial in aiding humans to detect machine-generated texts.

3. Design

CoAT is composed of 246k human-written and artificial texts. The corpus creation methodology includes three main stages: (i) collecting human-written texts, (ii) artificial text generation, and (iii) post-processing and filtering.

3.1 Human-written text collection

We collect human-written data from six domains that cover normative Russian, general domain texts, social media posts, texts from various historical periods, bureaucratic texts with complex discourse structure and embedded named entities, and other domains specified in the task-specific datasets, such as subtitles and web-texts. It is important to note that, in addition to linguistic and stylometric features, texts vary in length (e.g., sentence-level vs. document-level) and peculiarities inherent to the downstream tasks described in more detail below in Section 3.2.

Russian National Corpus. We use the diachronic sub-corpora of the Russian National Corpus (RNC),[g] which covers three historical periods and represents the modern Russian language ("pre-Soviet," "Soviet," and "post-Soviet").

Social media. We collect posts published between 2010 and 2020 using the X (Twitter) API by querying generic, frequently used hashtags. These hashtags are not tied to specific domains or topics and include terms such as days of the week, months, seasons, holidays, or the names of the

[g]The Russian National Corpus is a representative collection of texts in Russian, counting more than 1,5B tokens and completed with linguistic annotation and search tools. Available at ruscorpora.ru/en

https://doi.org/10.1017/nlp.2024.38 Published online by Cambridge University Press

most popular cities in Russia. These texts are generally short, written in an informal style, and may include emojis and obscene lexis. We anonymise the texts by excluding the user handles and IDs.
Wikipedia. We use paragraphs from the top 100 most viewed Russian Wikipedia pages spanning the period of 2016–2021 according to the PageViews[h] statistics.
News articles. The news segment covers different news sources in the Taiga corpus (Shavrina and Shapovalova 2017) and the `corus` library.[i] Since these corpora are publicly available, we additionally parse more recent news articles from various Russian news websites published at the end of 2021 to prevent potential data leakage so that the test set, at the moment of creating the corpus, does not contain samples that could not be retrieved from the news websites.
Digitalised personal diaries. We use texts from the Prozhito corpus (Melnichenko and Tyshkevich 2017), which includes digitilised personal diaries written during the 20th century.
Strategic documents. Here, we use strategic documents from the Ministry of Economic Development of the Russian Federation (Ivanin *et al.* 2020). The documents are written in a bureaucratic style, rich in embedded entities, and have complex syntactic and discourse structure.
Task-specific datasets. We collect gold standard references from the Wikimatrix (Schwenk *et al.* 2021) and Tatoeba (Tiedemann 2012) machine translation datasets since they are generally written and/or validated by human annotators (Artetxe and Schwenk 2019; Scialom *et al.* 2020; Hasan *et al.* 2021). To ensure that low-quality instances are not included in CoAT, we filter these datasets based on the sentence length and remove duplicates before translating them into Russian.

3.2 Artificial text generation

We use human-written texts as the input to 13 TGMs varying in their number of parameters, architecture choices, and pre-training objectives. Each model is fine-tuned for one or more of the following natural language generation tasks: machine translation, paraphrase generation, text simplification, and text summarisation. In addition, we consider back-translation and zero-shot generation approaches.
Machine translation & back-translation. We use three machine translation models via the `EasyNMT` framework:[j] OPUS-machine translation (Tiedemann and Thottingal 2020), M-BART50 (Tang *et al.* 2020), and M2M-100 (Fan *et al.* 2021). We select these models for their near state-of-the-art performance in English-Russian translation, the ease of use of the EasyNMT framework, and the diversity they offer in machine translation approaches OPUS-MT for one-to-one, M-BART50 and M2M-100 for many-to-many translations, with M-BART50 featuring a pretrained backbone. We use subsets of the Tatoeba (Artetxe and Schwenk 2019) and WikiMatrix (Schwenk *et al.* 2021) datasets to obtain translations among three language pairs: English-Russian, French-Russian, and Spanish-Russian. In the back-translation setting, the input sentence is translated into one of the target languages, and then back into Russian.
Paraphrase generation. Paraphrases are generated with models available under the `russian-paraphrasers` library (Fenogenova 2021a): ruGPT2-Large,[k] ruGPT3-Large (Zmitrovich *et al.* 2024), ruT5-Base-Multitask,[l] and mT5 (Xue *et al.* 2021) of Small and Large versions.
Text simplification. We fine-tune ruGPT3-Small (Zmitrovich *et al.* 2024), ruGPT3-Medium (Zmitrovich *et al.* 2024), ruGPT3-Large, mT5-Large, and ruT5-Large (Zmitrovich *et al.* 2024) for text simplification on a filtered version of the RuSimpleSentEval-2022 dataset (Sakhovskiy *et al.*

[h] pageviews.wmcloud.org
[i] github.com/natasha/corus
[j] github.com/UKPLab/EasyNMT
[k] hf.co/ai-forever/rugpt2large
[l] hf.co/cointegrated/rut5-base-multitask

https://doi.org/10.1017/nlp.2024.38 Published online by Cambridge University Press

2021; Fenogenova 2021b). Fine-tuning of each model is run for 4 epochs with a batch size of 4, learning rate of 10^{-5}, and weight decay of 10^{-2}.

Text summarisation. We use two abstractive summarisation models fine-tuned on the Gazeta dataset (Gusev 2020): ruT5-base[m] and M-BART.[n]

Open-ended generation. We generate texts in a zero-shot manner by prompting the model with a beginning of the text and specifying the maximum number of 500 tokens in the model output. The models include ruGPT3-Small, ruGPT3-Medium, ruGPT3-Large.

3.3 Post-processing and filtering

Each generated text undergoes a post-processing and filtering procedure based on a combination of language processing tools and heuristics. First, we discard text duplicates, copied inputs, and empty outputs, and remove special tokens (e.g., <s>, </s>, <pad>, etc.). Next, we discard texts containing obscene lexis according to the corpus of Russian obscene words.[o] We keep translations classified as Russian by the language detection model[p] with a confidence of more than 90%. Finally, we empirically define text length intervals for each natural language generation task based on a manual analysis of length distributions in `razdel`[q] this library tokens. The texts are filtered by the following token ranges: 5-to-25 (*machine translation, back-translation, paraphrase generation*), 10-to-30 (*text simplification*), 15-to-60 (*text summarisation*), and 85-to-400 (*open-ended generation*). Thus, we remove the possibility of using length as a feature to distinguish texts and reduce the size of the corpus by approximately 30%.

4. Corpus analysis

4.1 General statistics

Number of texts. Tables 2 and 3 summarise the distribution of texts in CoAT by natural language task, model, and domain. The number of human-written and machine-generated texts is balanced within each natural language task and domain. Texts from the Russian National Corpus are the most common in CoAT (20%). News articles make up a percentage of 19.6% of the total number of texts, followed by strategic documents (18.5%) and texts from Wikipedia (17.5%). Digitalised personal diaries comprise 10%, while texts from social media and machine translation datasets account for 7% and 6%, respectively.

Length and frequency. Table 4 presents a statistical analysis of CoAT based on the frequency and lexical richness metrics. We compute the token frequency in each text as the number of frequently used tokens (i.e., the number of instances per million in the Russian National Corpus is higher than one) divided by the total number of tokens in the text. The average text length in `razdel` tokens is similar for the sentence-level natural language generation tasks (*back-translation*, *machine translation*, and *paraphrase generation*), while texts produced by *text simplification* and *summarisation* models are of 20.59 and 31.45 tokens on average, respectively. The overall average text length is 49.64. The distribution of high-frequency tokens in human-written and machine-generated texts is similar within each generation task, comprising of 88% high-frequency tokens on average.

Lexical richness. We evaluate the lexical richness of the CoAT texts using three measures from the `lexicalrichness`[r] library: word count, terms count, and corrected type-token ratio (CTTR).

[m] hf.co/IlyaGusev/rut5-base-sum-gazeta
[n] hf.co/IlyaGusev/mbart-ru-sum-gazeta
[o] github.com/odaykhovskaya/obscene-words
[p] github.com/fedelopez77/langdetect
[q] github.com/natasha/razdel
[r] github.com/LSYS/LexicalRichness

https://doi.org/10.1017/nlp.2024.38 Published online by Cambridge University Press

https://doi.org/10.1017/nlp.2024.38 Published online by Cambridge University Press

Table 2. General statistics of CoAT by natural language generation task and model

Task	Model	# texts	Domain	Task	Model	# texts	Domain
Back-translation	Human	8,946	Russian National Corpus,	Machine	Human	7,026	WikiMatrix,
	M-BART50	2,816	Wikipedia, news articles,	translation	M-BART50	2,385	Tatoeba
	M2M-100	3,063	diaries, strategic documents,		M2M-100	2,331	
	OPUS-MT	3,067	WikiMatrix, Tatoeba		OPUS-MT	2,310	
Open-ended generation	Human	45,367	Russian National Corpus,	Text	Human	19,009	Russian National
	ruGPT3-Large	16,210	social media, Wikipedia,	summarisation	M-BART	3,549	Corpus, Wikipedia,
	ruGPT3-Medium	14,470	news articles, diaries,		M-BART50	5,788	news articles, diaries,
	ruGPT3-Small	14,687	strategic documents		ruT5-Base	9,672	strategic documents
Paraphrase generation	Human	22,838	Russian National Corpus,	Text	Human	20,055	Russian National Corpus,
	mT5-Large	2,516	social media, Wikipedia,	simplification	mT5-Large	2,805	social media, Wikipedia,
	mT5-Small	5,676	news articles, diaries,		ruGPT3-Small	4,305	news articles, diaries,
	ruGPT2-Large	5,626	strategic documents		ruGPT3-Medium	4,364	strategic documents
	ruGPT3-Large	3,590			ruGPT3-Large	4,239	
	ruT5-Base	5,430			ruT5-Large	4,342	

Table 3. Number of texts in each domain and task-specific dataset in CoAT

	Strategic documents	News articles	Digitalised diaries	Russian National Corpus	Social media	Tatoeba	WikiMatrix	Wikipedia
Back-translation	3,680	3,998	2,672	3,352	✗	784	1,650	1,756
Machine translation	✗	✗	✗	✗	✗	2,424	11,628	✗
Open-ended generation	17,154	19,288	3,026	23,412	7,312	✗	✗	20,542
Paraphrase generation	9,042	9,684	6,160	7,554	5,598	✗	✗	7,638
Text simplification	7,872	7,768	5,308	8,056	4,458	✗	✗	6,648
Text summarisation	7,916	7,736	7,568	8,034	✗	✗	✗	6,764
Overall	45,664	48,474	24,734	50,408	17,368	3,208	13,278	43,348

Table 4. Lexical richness metrics by natural language generation task in CoAT. **Notations:** %=the average fraction of high-frequency tokens. **H**=human-written texts. **M**=machine-generated texts

		%		Words		Terms		CTTR	
	Avg. num tokens	H	M	H	M	H	M	H	M
Back-translation	16.23	90.0	91.0	13.02	13.34	12.49	12.50	2.43	2.40
Machine translation	15.37	90.0	90.0	12.42	12.46	11.90	11.76	2.37	2.34
Open-ended generation	99.07	87.0	88.0	64.80	90.26	47.74	76.30	4.02	5.36
Paraphrase generation	15.73	88.0	88.0	13.08	12.91	12.54	12.54	2.43	2.45
Text simplification	20.59	88.0	89.0	16.51	16.55	15.40	15.05	2.64	2.58
Text summarisation	31.45	88.0	87.0	23.79	27.16	21.64	22.31	3.09	3.01
Overall	49.64	88.0	88.0	34.29	44.18	27.33	37.88	3.15	3.62

The corrected type-token ratio is calculated as $t/\sqrt{2 * w}$, where w is the total number of all words, including functional, and t is the number of unique terms in the vocabulary. We observe that the ratio of the measures between the natural and artificial texts depends on the task. In contrast to other natural language generation tasks, neural texts generated using the *open-ended generation* approach are generally longer than the human-written ones, and receive higher richness metrics. The reason is that the models try to generate texts up to the maximum number of 500 tokens, and may produce non-existent words, degenerated textual segments, or rare words.

4.2 Linguistic analysis

Stylometric features. Stylometry helps characterise the author's style based on various linguistic features, which are commonly applied in the ATD and human and neural AA tasks (He *et al.* 2004; Lagutina *et al.* 2019; Uchendu *et al.* 2019, 2020). This section aims to analyse the stylometric properties of human-written and machine-generated texts.

Following the motivation by Fröhling and Zubiaga (2021), we manually inspect a subset of the artificial texts to define stylometric features that potentially capture text generation errors (see Table 5). The features can be informally categorised as follows:

https://doi.org/10.1017/nlp.2024.38 Published online by Cambridge University Press

Table 5. Manually picked examples of text generation errors from the train set. The examples are automatically translated into English for illustration purposes

Error	Example
Inappropriate punctuation	"Окончил гимназию при Московском училище живописи., скульптуры.." He graduated from the gymnasium at the Moscow School of Painting., Sculpture.
Non-existent words	"Кошка, согласно исламским правилам, считается истинным животным в иснасле." A cat, according to Islamic rules, is considered a true animal in isnasl.
Repetitions	"Что вы делаете здесь, что вы делаете здесь, что вы делаете." What are you doing here, what are you doing here, what are you doing.
Abrupt ending	"В ходе проверки установлено наличие нарушений при" During the inspection, the presence of violations in the
Nonsensical texts	"Я хочу быть капусту, я хожу по сторонам." I want to be a cabbage, I walk around.
Decoding confusions	"<...> представитель пресс-службы ведомства Франсуа Бийо де Вильезасвйяуиашфмргпнб" < . . . > the representative of the press service of the department Francois Billou de Villezasviyauiashfmrgpnb
Hallucination	"There is evidence, что в начале XX века театр возвращается <...>" There is evidence that at the beginning of the XX century the theater returned < . . . >
Diverse errors	", бывало: совесть заначит мучить, сны не спите, богу богу не молишься." it used to happen: conscience mean to torment, dreams do not sleep, you do not pray to God God.

(1) **Surface and frequency features:** (i) the text's length in characters (*Length*), (ii) the fraction of punctuation marks in the text (*PUNCT*), (iii) the fraction of Cyrillic symbols in the text (*Cyrillic*), and (iv) *IPM*.

(2) **Count-based morphological features:** the fraction of (i) prepositions (*PREP*) and (ii) conjunctions in the text. The features are computed with `pymorphy2` (Korobov 2015), a morphological analyser for Russian and Ukrainian.

(3) **Readability measures:** (i) Flesch-Kincaid Grade (*FKG*), (ii) Flesch Reading-Ease (*FRE*), and (iii) LIX adapted to the Russian language. We use the `ruTS` library (Shkarin 2023) to compute the features.

Statistical analysis. Table 6 presents the summary of the stylometric feature values for the human-written and machine-generated texts and each text generator independently.[s] The *Length* is attributed to the natural language generation task; texts produced by MT and paraphrase generation models will be naturally shorter. In contrast, ruGPT3-Small, ruGPT3-Medium, and ruGPT3-Large generate longer texts, which is the case for the zero-shot generation setup. At the same time, *Cyrillic* varies between the models and can be specific to text translations, decoding confusions, copying parts of the input in another language, and hallucinations. We observe that the percentage of high-frequent words (*IPM*) in the texts is similar among the generators. The morphological and readability features differ on average between the TGMs and humans and between the TGMs. The results are supported by the Mann-Whitney U test used to compare the differences between the distributions of the stylometric features in human-written and machine-generated texts. There is a significant difference between the mean values for all features except for *IPM* and *PREP*, with $\alpha = 0.05$.

[s]We use the weighted subset of 32,200 texts for the linguistic analysis.

https://doi.org/10.1017/nlp.2024.38 Published online by Cambridge University Press

Table 6. The average values of the stylometric features in the corpus subset. Machine refers to the machine-generated texts

Model	Surface/Frequency				Morphology		Readability		
	Length	Cyrillic	PUNCT	IPM	CONJ	PREP	FKG	FRE	LIX
Human	250.209	0.808	0.032	0.883	2.216	4.531	11.286	32.371	72.582
Machine	299.942	0.811	0.027	0.885	1.857	3.86	9.606	36.981	69.18
ruGPT2-Large	97.987	0.835	0.018	0.869	0.512	2.137	8.673	35.550	70.331
ruGPT3-Small	639.683	0.774	0.028	0.879	4.540	9.530	10.337	43.367	67.428
ruGPT3-Medium	616.866	0.764	0.029	0.876	4.676	8.758	10.171	44.599	67.246
ruGPT3-Large	401.021	0.804	0.026	0.890	3.040	7.161	9.866	40.125	67.763
ruT5-Base	219.074	0.823	0.028	0.878	2.125	3.899	11.030	29.387	73.051
ruT5-Base-Multitask	91.795	0.824	0.029	0.891	0.843	1.588	8.095	37.471	66.645
ruT5-Large	181.967	0.820	0.030	0.889	1.937	3.192	13.402	23.434	77.657
mT5-Small	93.417	0.829	0.026	0.885	0.645	2.038	8.088	39.085	67.402
mT5-Large	101.100	0.824	0.028	0.891	0.936	1.908	8.387	39.071	66.975
M-BART	223.297	0.831	0.027	0.867	1.799	4.251	12.591	21.970	78.286
M-BART50	133.318	0.791	0.032	0.878	1.176	2.592	8.730	41.758	68.171
OPUS-MT	93.816	0.815	0.031	0.908	0.984	1.612	7.660	42.977	63.869
M2M-100	95.911	0.810	0.031	0.912	0.929	1.644	7.859	41.965	64.596

Analysis of the feature space. We use the principal component analysis (PCA; Pearson 1901) on the stylometric features computed on the weighted corpus subset. Figure 1 illustrates the 2-dimensional distribution of the features by generative model. A large overlapped portion among the texts indicates that the stylometric features may not be useful in solving the ATD tasks.

Discussion. The linguistic analysis indicates that distributions of most of the considered stylometric features underlying the human-written and machine-generated texts are not the same (see Figure 2). However, a substantial portion of the texts overlaps, meaning that the properties of the texts are similar among the generators. We conclude, that due to this overlap, it might be challenging to distinguish between the human-written and machine-generated texts and identify the author of a given text by utilising only stylometric features.

5. Experiments

5.1 Method

Non-neural detectors. We use two non-neural text detectors via the `scikit-learn` library (Pedregosa *et al.* 2011): a data-agnostic classifier, referred to as "Majority," that predicts the most frequent class label in the training data, and a logistic regression classifier trained on TF-IDF features computed on word N-grams with the N-gram range $\in [1;3]$, denoted as "Linear".

Transformer-based detectors. We experiment with four monolingual and cross-lingual Transformer LMs: ruBERT-base (Zmitrovich *et al.* 2024; 178M), ruRoBERTa-large (Zmitrovich

https://doi.org/10.1017/nlp.2024.38 Published online by Cambridge University Press

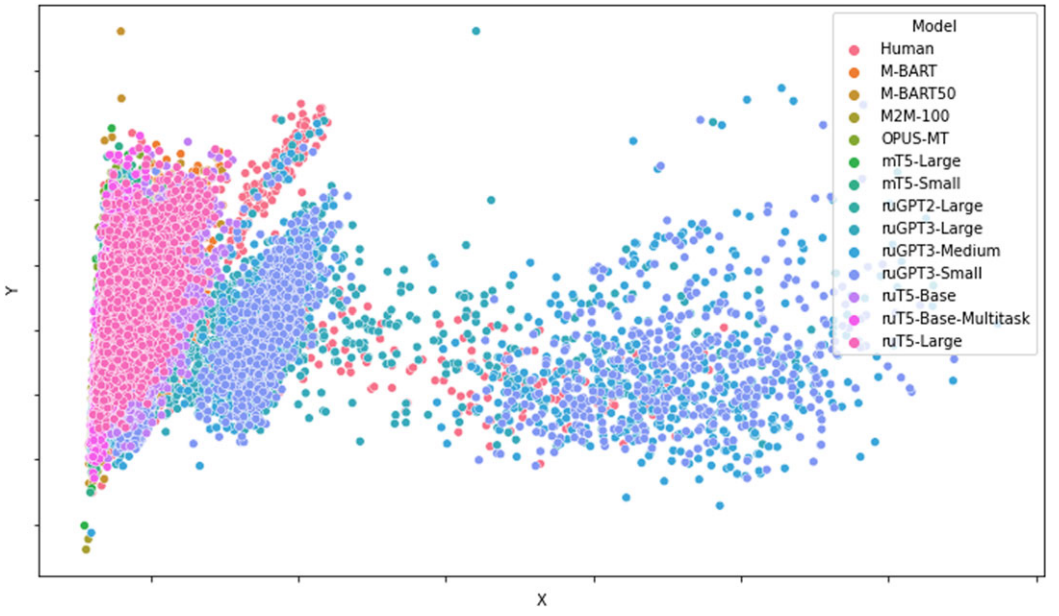

Figure 1. 2-dimensional distribution of the corpus subset using PCA.

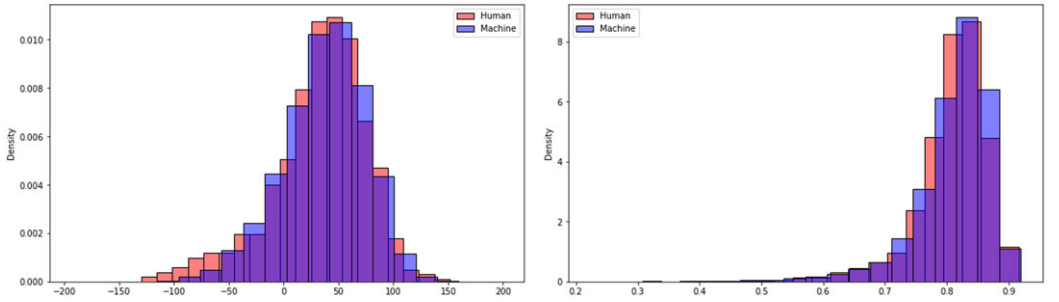

Figure 2. The distribution of the *FRE* (left) and *Cyrillic* (right) features between the human-written and machine-generated texts in the corpus subset. The difference between the mean values is statistically significant according to the Mann-Whitney U test, with the p-values equal to 0.002 and 0.0, respectively.

et al. 2024; 355M), XLM-R-base (Conneau *et al.* 2020; 278M), and RemBERT (Chung *et al.* 2020; 575M).

Performance metrics. The accuracy score is used as the target metric in the binary classification problems (Sections 5.2; 5.3; and 5.5), and macro-averaged F_1 is used as the target metric in the multi-class classification problem (Section 5.4). The results are averaged over three restarts with different random seeds.

Training details. We maximise the validation set performance by running a grid search over a set of hyperparameters. We tune the L_2 regularisation coefficient $C \in \{0.01, 0.1, 1.0\}$ to train the logistic regression model. We use the Transformer LMs' weights and codebase for fine-tuning and evaluation from the HuggingFace Transformers library (Wolf *et al.* 2020). The detectors are fine-tuned for 5 epochs over the learning rate 10^{-5}, with a fixed weight decay of 10^{-4} and batch sizes of 32 for RemBERT and 64 for the other LMs.

https://doi.org/10.1017/nlp.2024.38 Published online by Cambridge University Press

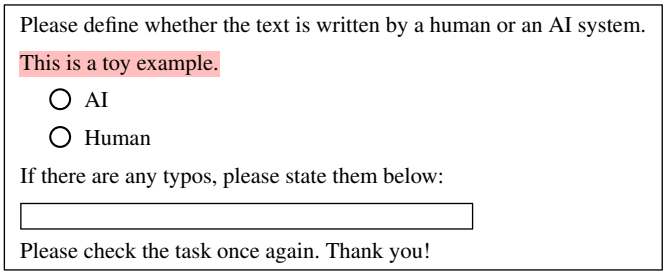

Figure 3. The web interface used for human evaluation on the artificial text detection task.

Splits. CoAT is split into train, validation, and private test sets in the 70/10/20 ratio in a stratified fashion (172k/24k/49k examples). We use the sets to create train, validation, and private test subsets for each experiment described below: (i) artificial text detection (Section 5.2), (ii) artificial text detection by natural language generation task (Section 5.3, (iii) authorship attribution (Section 5.4), (iv) robustness towards unseen TGM (Section 5.5.1), and (v) robustness towards unseen text domain (Section 5.5.2).

In Sections 5.2, 5.3, and 5.5, we train or fine-tune binary classifiers to distinguish between text written by a human and text generated with a TGM. In Section 5.4, the machine-generated label in the artificial text detection task is broken into 13 model names in the authorship attribution task. In each experiment configuration, we use the corresponding subsets of the CoAT train, validation, and private test set splits, which are balanced by the number of examples per target class, natural language generation task, TGM, and text domain.

5.2 Artificial text detection

Task formulation. This experiment aims to evaluate the performance of detectors in determining if a given text is automatically generated or written by a human. This task is framed as a binary classification problem with two labels: **H** (human-written) and **M** (machine-generated).

Splits. CoAT is split into train, validation, and private test sets in the 70/10/20 ratio in a stratified fashion (172k/24k/49k examples).

Human baseline. We conduct a human evaluation on the artificial text detection task using a stratifiedsubset of 5k samples from the test set. The evaluation is run via Toloka (Pavlichenko, Stelmakh, and Ustalov 2021), a crowd-sourcing platform for data labelling.[t] The annotation setup follows the conventional crowd-sourcing guidelines for the ATD task and accounts for methodological limitations discussed in Ippolito *et al.* (2020); Clark *et al.* (2021). We provide a full annotation instruction in Appendix A and an example of the Toloka web interface in Figure 3.

The human evaluation project consists of an unpaid training stage and the main annotation stage with honeypot task for tracking annotators' performance. The honeypot tasks are manually picked quality verification examples from the CoAT training set, which are mixed in with the main unlabelled examples. We compute the annotation performance by comparing the annotators' votes with the honeypot examples' gold labels. Before starting, the annotator receives a detailed instruction describing the task and showing annotation examples. The instruction is available anytime during the training and main annotation stages. Access to the project is granted to annotators ranked top-70% according to the Toloka rating system. Each annotator must finish the training stage by answering at least 27 out of 32 examples correctly. Each of the trained annotators gets paid. The pay rate is on average $2.55/h, which is twice the amount of the hourly minimum wage

[t] toloka.ai

https://doi.org/10.1017/nlp.2024.38 Published online by Cambridge University Press

Table 7. Accuracy scores of the detectors on the artificial text detection task

Majority	Linear	ruBERT	ruRoBERTa	RemBERT	XLM-RoBERTa	Human
0.50	0.733	0.856	**0.876**	0.866	0.859	0.66

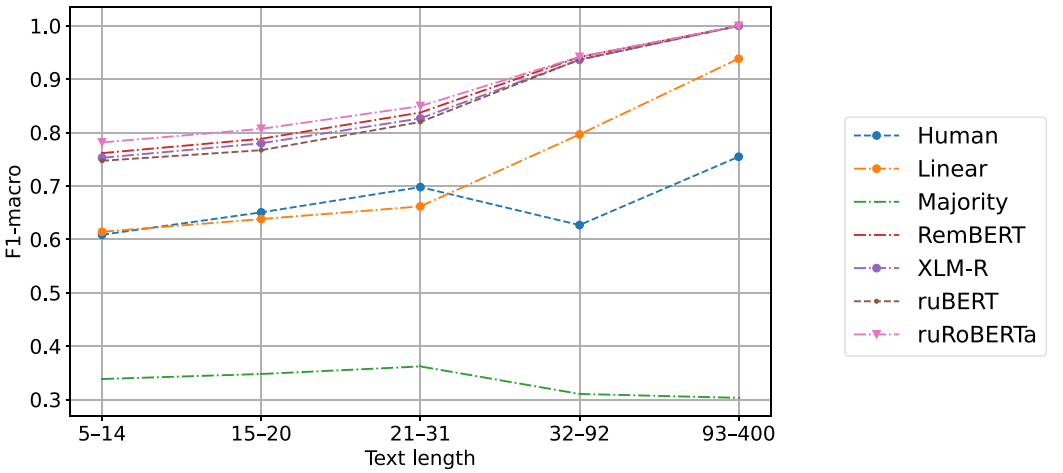

Figure 4. Macro-F1 scores of the detectors on the artificial text detection task grouped by five quintiles of the text length.

in Russia. We aggregate the majority vote labels via the dynamic overlap from three to five trained annotators after (i) discarding votes from annotators whose annotation performance on the honeypot tasks is less than 50% and (ii) filtering out submissions with less than 15 s of response time per five texts.

Results. Table 7 presents the results of this experiment. Transformer-based detectors outperform non-neural detectors by a wide margin with an average accuracy of 0.864. The highest accuracy scores were achieved using ruRoBERTa, followed by RemBERT, ruBERT, and XLM-RoBERTa. Notably, using monolingual ruRoBERTa and ruBERT led to a significant improvement in accuracy scores, despite having fewer parameters than XLM-RoBERTa and RemBERT. At the same time, human annotators significantly underperform the detectors. The human evaluation yields an overall accuracy of 0.66, which is 0.07 points lower than that of non-neural detectors and 0.2 points lower than that of neural detectors.

Effect of length. We divide the test set into five groups of equal size based on the text length to examine the effect of text length on the performance. As depicted in Figure 4, the F1-macro-scores of learnable detectors improved monotonically as the text length increases. For the shortest texts, non-neural and neural detectors achieved F1-macro-scores of 0.6 and 0.75, respectively. The scores increase significantly, resulting in near-perfect predictions of 0.95 to 0.99 for the longest texts. The differences between the four Transformer-based detectors are most prominent in shorter texts but became less significant as the text length increases.

Humans exhibit comparable or better performance than non-neural detectors when it comes to texts that are up to 31 words. However, humans seem to face difficulty in classifying texts that fall within the range of 32–92 words, while performing best in identifying the longest texts. A possible explanation for this consistent performance of humans within the 0.6 to 0.75 range of the macro-F1 could be that humans tend to rely more on surface-level features, which may have a similar distribution across all text length groups.

https://doi.org/10.1017/nlp.2024.38 Published online by Cambridge University Press

Discussion. The results demonstrate that state-of-the-art Transformer-based models can be relatively successful in distinguishing human-written texts from machine-generated ones for the Russian language. However, one can quickly notice a rather stark contrast between the best scores obtained on the CoAT test set in binary setup (0.86 accuracy) and scores obtained for a similar setup in English (0.970 accuracy; see Uchendu *et al.* (2020) for reference). We attribute this disparity not to the language discrepancy but primarily to the nature of the texts. In the English setup, the average text length was 432 words, which is nearly nine times longer than the CoAT's 50-word average. Our findings align with the works by Munir *et al.* (2021); Stiff and Johansson (2022), which report that shorter texts are the most challenging for both human annotators and computational methods.

The low results in human evaluation are consistent with recent works by Karpinska *et al.* (2021); Uchendu *et al.* (2021), which underpin the difficulty of the generated text detection task for crowd-sourcing annotators. These works advise hiring experts trained to evaluate written texts or conduct multiple crowd-sourcing evaluation setups with extensive training stages.

5.3 Artificial text detection by task

Task formulation. In this experiment, we create six datasets with respect to each natural language generation task. The detectors are independently trained and evaluated on the human-written and machine-generated texts from the same task. This setup allows for estimating the complexity of the tasks. The lower the performance scores, the more natural the TGMs in a particular task.

Splits. CoAT is split into train, validation, and private test sets in the 70/10/20 ratio in a stratified fashion: 9.8k/1.4k/2.7k (*machine translation*), 12.4k/1.7k/3.4k (*back-translation*), 31.8k/4.4k/8.9k (*paraphrase generation*), 26.6k/3.7k/7.3k (*text summarisation*), 28k/3.8k/7.9k (*text simplification*), and 63k/9.2k/17.9k (*open-ended generation*).

Results. The results of our experiment are presented in Table 8. Averaged macro-F_1 scores are presented in Table 9. We observe that the detectors' performance rankings are consistent with their performance in the artificial text detection setup (Section 5.2). ruRoBERTa consistently outperforms the other detectors with an average macro-F_1 of 0.814, while other models show moderate performance. Analysing the results by a natural language generation task, we find that machine-translated texts are more difficult to detect, where the best detector performs only slightly better than random prediction, with an accuracy of 0.621 and macro-F_1 of 0.612. Tasks such as **Back-translation**, **Paraphrase Generation**, and **Text Simplification** are of intermediate difficulty, resulting in an accuracy of 0.768 and macro-F_1 of 0.765 or above. Finally, we find that **Text Summarisation** and **Open-ended Generation** are much easier to detect. In fact, the detectors made near-perfect predictions in **Open-ended Generation**, with accuracies and macro-F_1 scores reaching up to 0.99.

Discussion. We suggest that the diversity of TGMs' output and the degree of control over the outputs in the task affect the task difficulty. The TGM outputs of **Machine Translation** and **Paraphrase Generation** are semantically constrained by the inputs, thus it is less likely for these TGM to produce in-plausible or repetitive outputs. **Open-ended Generation** models suffer from hallucination and repetitions, which at scale can be learnt by detectors. At the same time.

5.4 Authorship attribution

Task formulation. The author attribution task aims at determining the author of a given text. The task is framed as a multi-class classification problem with 14 target classes: a human author and 13 TGMs. In this experiment, we use the same dataset as in Section 5.2, but this time instead of binary labels we use the source TGM's labels as prediction targets. In this setup, the dataset is

https://doi.org/10.1017/nlp.2024.38 Published online by Cambridge University Press

Table 8. Macro-F_1 and accuracy scores by a natural language generation task

Task	Model	Macro-F_1	Acc.	Task	Model	Macro-F_1	Acc.
Back-translation	Majority	0.333	0.500	Paraphrase generation	Majority	0.333	0.500
	Linear	0.664	0.665		Linear	0.677	0.677
	ruBERT	0.792	0.792		ruBERT	0.823	0.824
	ruRoBERTa	0.820	0.821		ruRoBERTa	0.844	0.845
	RemBERT	0.792	0.792		RemBERT	0.822	0.822
	XLM-RoBERTa	0.765	0.768		XLM-RoBERTa	0.815	0.816
Machine translation	Majority	0.333	0.500	Text simplification	Majority	0.333	0.500
	Linear	0.557	0.557		Linear	0.660	0.660
	ruBERT	0.629	0.630		ruBERT	0.801	0.801
	ruRoBERTa	0.675	0.677		ruRoBERTa	0.847	0.847
	RemBERT	0.641	0.648		RemBERT	0.641	0.648
	XLM-RoBERTa	0.612	0.621		XLM-RoBERTa	0.816	0.816
Open-ended generation	Majority	0.333	0.500	Text summarisation	Majority	0.333	0.500
	Linear	0.937	0.938		Linear	0.702	0.703
	ruBERT	0.997	0.997		ruBERT	0.862	0.862
	ruRoBERTa	0.998	0.998		ruRoBERTa	0.884	0.884
	RemBERT	0.997	0.997		RemBERT	0.880	0.880
	XLM-RoBERTa	0.996	0.996		XLM-RoBERTa	0.870	0.869

Table 9. Averaged macro-F_1 by a natural language generation task

	Majority	Linear	ruBERT	ruRoBERTa	RemBERT	XLM-RoBERTa
Avg. macro-F_1	0.333	0.700	0.817	0.845	0.796	0.812

imbalanced: 50% of samples are human-written and the other 50% of samples contain outputs of 13 TGMs. In this case, we rely on macro-F_1 as the main performance metric.

Splits. CoAT is split into train, validation, and private test sets in the 70/10/20 ratio in a stratified fashion (172k/24k/49k examples).

Results. Our findings, shown in Table 10, demonstrate that RuRoBERTa significantly outperforms other models with a macro-F_1 score of 0.521. RemBERT and ruBERT performed similarly with macro-F_1 scores of 0.496 and 0.476, while XLM-RoBERTa achieved a lower macro-F_1 score of 0.451. These findings align with the results of artificial text detection (Section 5.2) and artificial text detection by task (Section 5.3). In terms of TGMs, the neural detectors achieved the highest average performance of 0.75 F_1 in detecting ruT5-Base, while the lowest average performance of 0.112 F_1 was observed in detecting ruT5-Large. However, the TGM size did not significantly affect the performance of neural detectors in the ruGPT3 family, as the difference in F_1 between detecting ruGPT3-Small and ruGPT3-Large was only 0.01. In summary, neural detectors exhibit significantly higher accuracy in identifying human-written texts (with an average F_1 score of 0.866) compared to determining the source TGM.

Discussion. Authorship attribution is important when legal requirements demand revealing text generation models for transparency, claiming intellectual property, or replicating text generation. The task of authorship attribution is more challenging than the binary artificial text detection task,

https://doi.org/10.1017/nlp.2024.38 Published online by Cambridge University Press

Table 10. F_1 scores of the authorship attribution task by the target TGM

TGM	Majority	Linear	ruBERT	ruRoBERTa	RemBERT	XLM-RoBERTa
Human	0.667	0.749	0.860	0.882	0.868	0.856
ruGPT2-Large	0.0	0.23	0.652	0.675	0.685	0.652
ruGPT3-Small	0.0	0.48	0.651	0.778	0.677	0.652
ruGPT3-Medium	0.0	0.39	0.545	0.714	0.614	0.440
ruGPT3-Large	0.0	0.43	0.643	0.724	0.688	0.647
ruT5-Base	0.0	0.23	0.746	0.758	0.758	0.741
ruT5-Base-Multitask	0.0	0.0	0.217	0.231	0.218	0.158
ruT5-Large	0.0	0.0	0.122	0.121	0.122	0.083
mT5-Small	0.0	0.08	0.504	0.536	0.499	0.461
mT5-Large	0.0	0.0	0.229	0.243	0.252	0.138
M-BART	0.0	0.10	0.506	0.533	0.540	0.522
M-BART50	0.0	0.23	0.476	0.501	0.502	0.480
OPUS-MT	0.0	0.02	0.288	0.332	0.298	0.296
M2M-100	0.0	0.0	0.228	0.265	0.222	0.185
Macro-F_1	0.05	0.21	0.476	0.521	0.496	0.451

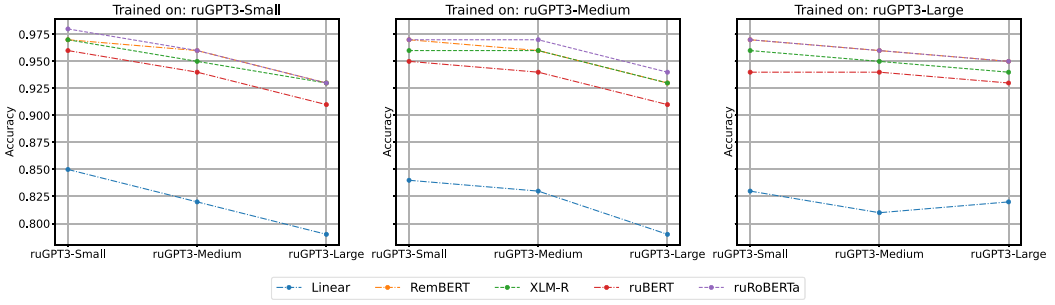

Figure 5. Results of testing the detectors' robustness towards the size of unseen GPT-based text generation models.

as observed in recent works (Uchendu *et al*. 2020, 2021). The low-performance scores indicate that it is challenging to differentiate between the TGMs, as they may lack distinct features that set them apart from each other, which might align with the results of the corpus linguistic analysis in Section 4.

5.5 Robustness

5.5.1 Size of unseen GPT-based text generaton models

Task formulation. This experiment setting tests the detectors' robustness towards the size of unseen GPT-based TGMs. The detectors are trained or fine-tuned on the human-written texts and

https://doi.org/10.1017/nlp.2024.38 Published online by Cambridge University Press

texts generated by one of the GPT-based models (ruGPT3-Small, ruGPT3-Medium, and ruGPT3-Large) as described in Section 5.1 and further inferred to detect texts from the out-of-domain GPT-based TGMs, i.e., those that are not present the train set. Consider an example for the ruGPT3-Small model, where we train the detectors on a mixture of human-written texts and texts generated by *only* ruGPT3-Small, and evaluate the detectors on three mixtures of human-written texts and texts generated by ruGPT3-Small, ruGPT3-Medium, and ruGPT3-Large.

Splits. CoAT is split into train, validation, and private test sets in the 70/10/20 ratio in a stratified fashion: 26.2k/3.7k/7.7k (ruGPT3-small), 26k/3.7k/7.6k (ruGPT3-medium), and 28.4k/4k/8.2k (ruGPT3-large).

Results. The results are presented in Figure 5. We observe that larger models' outputs are more challenging to identify as all figures show an evident drop in performance. However, training on larger models' outputs improves the detectors' robustness, as lines in the left-most plot become more flattened. Analysing the results by the detector, we find that all neural detectors achieve a strong performance of more than 90% accuracy. ruRoBERTa performs on par with RemBERT while having fewer parameters, and the linear detector receives moderate performance falling behind the Transformers by a large margin.

Discussion. Overall, the results align with the works of Solaiman *et al.* (2019) and Kushnareva *et al.* (2021). The ATD task can become more challenging with scaled TGMs since the TGMs' size directly determines the detector performance. However, the detectors receive the most optimal performance concerning different TGMs' sizes when trained on outputs from the largest ones.

5.5.2 Unseen text domain

Task formulation. Here, we analyse whether the detectors are robust in classifying TGMs' outputs from unseen text domains. Similar to Section 5.5.1, we train or fine-tune the detectors on the human-written and machine-generated texts from one domain and evaluate them on the texts from the other domains. Consider an example for the news domain, where we train the detectors on a mixture of human-written and machine-generated texts from *only* the news domain, and evaluate the detectors on six mixtures of human-written and machine-generated texts from the domains of RNC, social media, Wikipedia, digitalised diaries (Prozhito), strategic documents (Minek), and news.

Splits. CoAT is split into train, validation, and private test sets in the 70/10/20 ratio in a stratified fashion: 30.8k/4.4k/10.2k (strategic documents), 34.7k/4.9k/8.5k (news articles), 17.2k/2.4k/4.9k (digitalised diaries), 36.3k/5.2k/8.5k (Russian National Corpus), 12.5k/1.7k/2.9k (social media), and 28.8k/4k/9.9k (Wikipedia).

Results. Figure 6 presents the results of this experiment. Overall, the detectors demonstrate similar transferred performance across all domains. The in-domain performance is up to 90% regardless of the detectors' number of weight parameters and architecture. The detection is more reliable when training on texts from News and Wikipedia as indicated by less spiky patterns in the corresponding figures showing results for the models trained on these subsets. However, training on texts from Minek, Prozhito, and Social Media may result in near-random performance on the out-of-domain test sets, as corresponding figures show a single outstanding peak. We also observe that ruRoBERTa receives higher transferred accuracy than the other detectors, as seen from the purple dash-dotted line being on top of other lines in almost all patterns.

Discussion. While the TGMs are rapidly proliferating in different areas of life, the related ATD research primarily focuses on one particular domain (see Section 2.1). The single-domain evaluation limits the analysis of the detectors' limitations. Several works report that existing detectors exhibit inconsistent multi-domain performance (Bakhtin *et al.* 2019; Kushnareva *et al.* 2021). To our knowledge, our work is among the first to analyse the detectors' cross-domain generalisation. We empirically show that the detectors fail to transfer when trained on specific domains, such as strategic financial documents and social media posts.

https://doi.org/10.1017/nlp.2024.38 Published online by Cambridge University Press

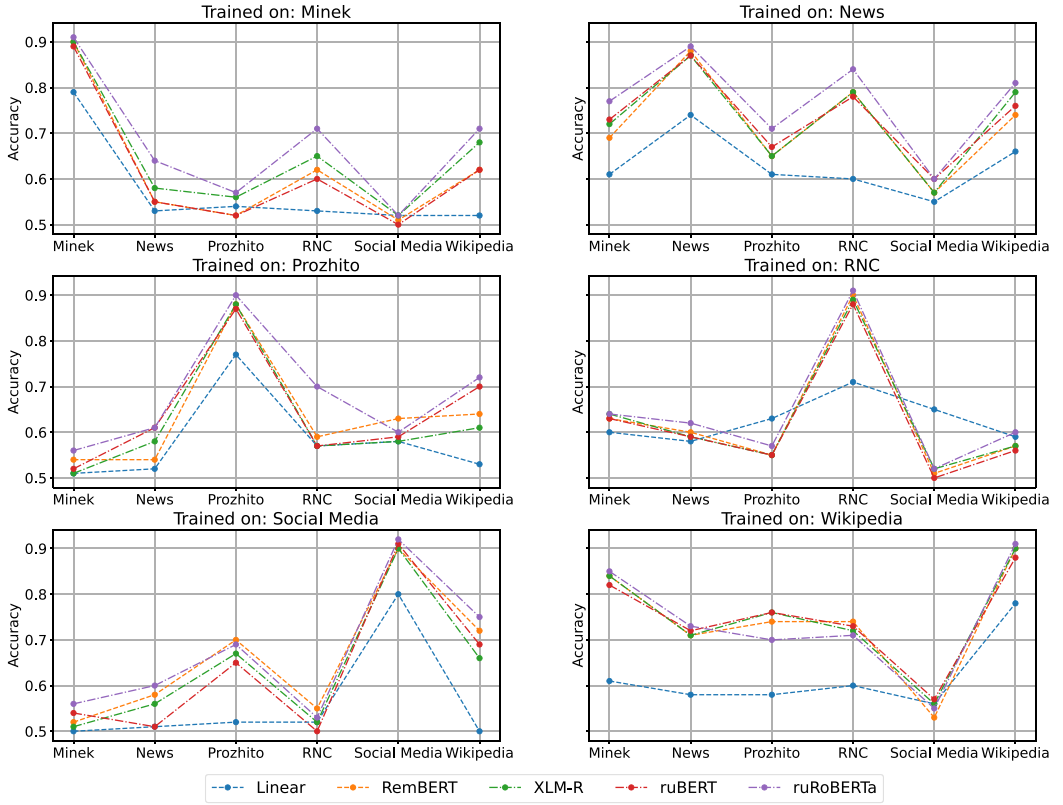

Figure 6. Results of testing the detectors' robustness towards unseen text domains. **Notations:** Minek=strategic documents; Prozhito=digitalised diaries; RNC = Russian National Corpus.

6. Conclusion and future work

This work proposes CoAT, a large-scale corpus for neural text detection in Russian. CoAT comprises more than 246k human-written and machine-generated texts, covering various generative language models, natural language generation tasks, and text domains. Our corpus promotes the development of multi-domain artificial text detectors to warn humans about potentially generated content on news and social media platforms, such as fake news, generated product reviews, and propaganda spread with bots. We present a linguistic analysis of our corpus and extensively evaluate the feature-based and Transformer-based artificial text detectors. The key empirical results indicate that humans struggle to detect the generated text. At the same time, the detectors fail to transfer when trained on outputs from smaller TGMs and specific text domains.

In this paper, we explore multiple experimental setups in which we find the following:

(i) fine-tuning state-of-the-art Transformer-based models to determine whether the text was written by a human or generated by a machine leads to satisfactory results but leaves room for further improvement.

(ii) it is more difficult to detect texts generated by conditioned text generation models compared to open-ended generation.

(iii) determining the source text generation model is more difficult than determining whether the text was machine-generated.

https://doi.org/10.1017/nlp.2024.38 Published online by Cambridge University Press

(iv) fine-tuned detectors are not robust towards the size of the text generation model. Larger models are more difficult to detect.

(v) fine-tuned detectors are not robust towards the unseen text domain.

These observations underscore the challenge of implementing a trustworthy detector in real-life applications, where there is no information available about the domain and potential text generation model.

In our future work, we aim to explore ATD tasks in the multilingual setting. Another direction is to analyse the effect of the human evaluation design on the performance, e.g., the varying number of training examples and providing the input texts in the sequence-to-sequence tasks.

7. Ethical considerations

Crowd-sourcing annotation. Responses of human annotators are collected and stored anonymously. The average annotation pay is double the hourly minimum wage in Russia. The annotators are warned about potentially sensitive topics in data (e.g., politics, culture, and religion).

Social and ethical risks. The scope of risks associated with the misuse of generative language models is widely discussed in the community (Weidinger *et al.* 2021; Bommasani *et al.* 2021). This problem has been addressed from the perspective of responsible artificial intelligence development: researchers and developers create new regulations and licenses (Contractor *et al.* 2022), require outputs from the TGMs to be marked as "generated," and propose "watermarking" techniques to determine generated content (Kirchenbauer *et al.* 2023). While our goal is to propose a novel large-scale ATD resource for the Russian language, we understand that the results of our work can be used maliciously, e.g., to reduce the performance of the detectors. However, we believe that CoAT will contribute to the development of more generalisable detectors for non-English languages.

8. Limitations

Data collection. Learnable artificial text detection methods require human-written and machine-generated texts. The design of the ATD resources is inherently limited by the availability of diverse text domains and generative language models. In particular, such resources may suffer from decreasing inclusivity of the models due to the rapid development of the field of natural language generation. While we have addressed the diversity of the corpus in terms of the TGMs and text domains, it is crucial to continuously update CoAT to keep it up-to-date with the recent TGMs and conduct additional evaluation experiments.

Decoding strategies. The choice of the decoding strategy affects the quality of generated texts (Ippolito *et al.* 2019) and the performance of artificial text detectors (Holtzman *et al.* 2020). The design of CoAT does not account for the diversity of decoding strategies, limiting the scope of the detectors' evaluation. We leave this aspect for future work.

Acknowledgments. We acknowledge the computational resources of HPC facilities at the HSE University. We thank Elena Tutubalina, Daniil Cherniavskii, and Ivan Smurov for their contribution to the project in the early stages. We also appreciate including the earlier CoAT version in the SemEval-2024 Task 8 corpus (Wang *et al.* 2024b) for developing generalisable multi-domain, multimodel, and multilingual detectors.

Competing interests. The author(s) declare none.

References

Abbasi A. and Chen H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)* **26**(2), 1–29.

Adelani D. I., Mai H., Fang F., Nguyen H. H., Yamagishi J. and Echizen I. (2020). Generating sentiment-preserving fake online reviews using neural language models and their human-and machine-based detection. In *Advanced Information Networking and Applications: Proceedings of the 34th International Conference on Advanced Information Networking and Applications (AINA-2020)*, Springer, pp. 1341–1354.

Aharoni R., Koppel M. and Goldberg Y. (2014). Automatic detection of machine translated text and translation quality estimation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Baltimore, Maryland: Association for Computational Linguistics, pp. 289–295.

Artetxe M. and Schwenk H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics* **7**, 597–610.

Badaskar S., Agarwal S. and Arora S. (2008). Identifying real or fake articles: Towards better language modeling. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Bahri D., Tay Y., Zheng C., Brunk C., Metzler D. and Tomkins A. (2021). Generative models are unsupervised predictors of page quality: A colossal-scale study. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 301–309.

Bakhtin A., Gross S., Ott M., Deng Y., Ranzato M. and Szlam A. (2019). Real or Fake? Learning to Discriminate Machine from Human Generated Text. *arXiv preprint arXiv:* 1906.

Bommasani R., Hudson D. A., Adeli E., Altman R., Arora S., von Arx S., Bernstein M. S., Bohg J., Bosselut A., Brunskill E. and et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:* 2108.07258.

Chen X., Jin P., Jing S. and Xie C. (2022). Automatic detection of Chinese generated essayss based on pre-trained bert. In *2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, IEEE, vol 10, pp. 2257–2260.

Chung H. W., Fevry T., Tsai H., Johnson M. and Ruder S. (2020). Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*

Clark E., August T., Serrano S., Haduong N., Gururangan S. and Smith N. A. (2021). All that's 'human' is not gold: Evaluating human evaluation of generated text. 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), Online*, Association for Computational Linguistics, pp. 7282–7296.

Clark K., Khandelwal U., Levy O. and Manning C. D. (2019). What does BERT look at? An analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Florence, Italy: Association for Computational Linguistics, pp. 276–286.

Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L. and Stoyanov V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, pp. 8440–8451.

Contractor D., McDuff D., Haines J. K., Lee J., Hines C., Hecht B., Vincent N. and Li H. (2022). Behavioral use licensing for responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 778–788.

Diwan N., Chakraborty T. and Shafiq Z. (2021). Fingerprinting fine-tuned language models in the wild. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP. 2021*. Online: Association for Computational Linguistics, pp. 4652–4664.

Dugan L., Ippolito D., Kirubarajan A. and Callison-Burch C. (2020). RoFT: A tool for evaluating human detection of machine-generated text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online: Association for Computational Linguistics, pp. 189–196.

Dugan L., Ippolito D., Kirubarajan A., Shi S. and Callison-Burch C. (2023). Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. In *Proceedings of the 2023 AAAI Conference on Artificial Intelligence*.

Fagni T., Falchi F., Gambini M., Martella A. and Tesconi M. (2021). TweepFake: about detecting deepfake tweets. *Plos One* **16**(5), e0251415.

Fan A., Bhosale S., Schwenk H., Ma Z., El-Kishky A., Goyal S., Baines M., Celebi O., Wenzek G., Chaudhary V. and et al. (2021). Beyond English-centric multilingual machine translation. *Journal of Machine Learning Research* **22**(107), 1–48.

Fenogenova A. (2021a). Russian paraphrasers: Paraphrase with transformers. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, Kiyv, Ukraine: Association for Computational Linguistics, pp. 11–19.

Fenogenova A. (2021b). Text simplification with autoregressive models.

Fröhling L. and Zubiaga A. (2021). Feature-based detection of automated language models: Tackling GPT-2, GPT-3 and GROVER. *PeerJ Computer Science* **7**, e443.

Gallé M., Rozen J., Kruszewski G. and Elsahar H. (2021). Unsupervised and distributional detection of machine-generated text. *arXiv preprint arXiv:* 2111.02878.

https://doi.org/10.1017/nlp.2024.38 Published online by Cambridge University Press

Gehrmann S., **Strobelt H. and Rush A.** (2019). GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Florence, Italy: Association for Computational Linguistics, pp. 111–116.

Gupta S., **Nguyen H. H.**, **Yamagishi J. and Echizen I.** (2020). Viable threat on news reading: Generating biased news using natural language models. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science, Online*, Association for Computational Linguistics, pp. 55–65.

Gusev I. (2020). Dataset for automatic summarization of Russian news. In *Artificial Intelligence and Natural Language*. Cham: Springer International Publishing, pp. 122–134.

Hasan T., **Bhattacharjee A.**, **Islam M. S.**, **Mubasshir K.**, **Li Y.-F.**, **Kang Y.-B.**, **Rahman M. S. and Shahriyar R.** (2021). XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP. 2021, Online*. Association for Computational Linguistics, pp. 4693–4703.

He R. C., **Rasheed K. and et al.** (2004). Using machine learning techniques for stylometry. In *IC-AI*, pp. 897–903.

Holmes D. I. (1994). Authorship attribution. *Computers and the Humanities* **28**(2), 87–106.

Holmes D. I. (1998). The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing* **13**(3), 111–117.

Holtzman A., **Buys J.**, **Du L.**, **Forbes M. and Choi Y.** (2020). The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Ippolito D., **Duckworth D.**, **Callison-Burch C. and Eck D.** (2020). Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, pp. 1808–1822.

Ippolito D., **Kriz R.**, **Sedoc J.**, **Kustikova M. and Callison-Burch C.** (2019). Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, pp. 3752–3762.

Ivanin V., **Artemova E.**, **Batura T.**, **Ivanov V.**, **Sarkisyan V.**, **Tutubalina E. and Smurov I.** (2020). Rurebus-2020 shared task: Russian relation extraction for business. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"]*, Moscow, Russia.

Jawahar G., **Abdul-Mageed M.**, **Lakshmanan V. S. and** L. (2020). Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics (Online)*, Barcelona, Spain: International Committee on Computational Linguistics, pp. 2296–2309.

Jones K., **Nurse J. R. and Li S.** (2022). Are you Robert or RoBERTa? Deceiving online authorship attribution models using neural text generators. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. **16**, pp. 429–440.

Karpinska M., **Akoury N. and Iyyer M.** (2021). The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 1265–1285.

Kashnitsky Y., **Herrmannova D.**, **de Waard A.**, **Tsatsaronis G.**, **Fennell C. C. and Labbe C.** (2022). Overview of the DAGPap22 shared task on detecting automatically generated scientific papers. In *Proceedings of the Third Workshop on Scholarly Document Processing*, Gyeongju, Republic of Korea: Association for Computational Linguistics, pp. 210–213.

Kirchenbauer J., **Geiping J.**, **Wen Y.**, **Katz J.**, **Miers I. and Goldstein T.** (2023). A watermark for large language models.

Korobov M. (2015). Morphological analyzer and generator for Russian and Ukrainian languages. In M. Y. Khachay, N. Konstantinova, A. Panchenko, D. I. Ignatov, & V. G. Labunets, (eds), *Analysis of Images, Social Networks and Texts*, Communications in Computer and Information Science, vol. **542**, Springer International Publishing, pp. 320–332.

Kushnareva L., **Cherniavskii D.**, **Mikhailov V.**, **Artemova E.**, **Barannikov S.**, **Bernstein A.**, **Piontkovskaya I.**, **Piontkovski D. and Burnaev E.** (2021). Artificial text detection via examining the topology of attention maps. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 635–649.

Lagutina K., **Lagutina N.**, **Boychuk E.**, **Vorontsova I.**, **Shliakhtina E.**, **Belyaeva O.**, **Paramonov I. and Demidov P.** (2019). A survey on stylometric text features. In *2019 25th Conference of Open Innovations Association (FRUCT)*, IEEE, pp. 184–195.

Liu Y., **Ott M.**, **Goyal N.**, **Du J.**, **Joshi M.**, **Chen D.**, **Levy O.**, **Lewis M.**, **Zettlemoyer L. and Stoyanov V.** (2019). RoBERTa: A robustly optimized BERT pretraining approach.

Liyanage V., **Buscaldi D. and Nazarenko A.** (2022). A benchmark corpus for the detection of automatically generated text in academic publications. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, pp. 4692–4700.

Manjavacas E., **De Gussem J.**, **Daelemans W. and Kestemont M.** (2017). Assessing the stylistic properties of neurally generated text in authorship attribution. In *Proceedings of the Workshop on Stylistic Variation*, Copenhagen, Denmark: Association for Computational Linguistics, pp. 116–125.

Melnichenko M. and Tyshkevich N. (2017). Prozhito from manuscript to corpus. *ISTORIYA* **8**(7(61)).

Mitchell E., **Lee Y.**, **Khazatsky A.**, **Manning C. D. and Finn C.** (2023, DetectGPT: Zero-shot machine-generated text detection using probability curvature.

https://doi.org/10.1017/nlp.2024.38 Published online by Cambridge University Press

Munir S., **Batool B.**, **Shafiq Z.**, **Srinivasan P. and Zaffar F.** (2021). Through the looking glass: Learning to attribute synthetic text generated by language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online: Association for Computational Linguistics, pp. 1811–1822.

Pavlichenko N., **Stelmakh I. and Ustalov D.** (2021). Crowdspeech and Vox DIY: Benchmark dataset for crowdsourced audio transcription. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, vol. **1**.

Pearson K. and LIII (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**(11), 559–572.

Pedregosa F., **Varoquaux G.**, **Gramfort A.**, **Michel V.**, **Thirion B.**, **Grisel O.**, **Blondel M.**, **Prettenhofer P.**, **Weiss R.**, **Dubourg V. and et al.** (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* **12**, 2825–2830.

Pillutla K., **Swayamdipta S.**, **Zellers R.**, **Thickstun J.**, **Welleck S.**, **Choi Y. and Harchaoui Z.** (2021). MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J.W. Vaughan, (eds), *Advances in Neural Information Processing Systems*, vol. **34**, Curran Associates, Inc., pp. 4816–4828.

Radford A., **Narasimhan K.**, **Salimans T. and Sutskever I.** (2018). Improving language understanding by generative pre-training.

Radford A., **Wu J.**, **Child R.**, **Luan D.**, **Amodei D. and Sutskever I.** (2019). Language models are unsupervised multitask learners.

Rodriguez J., **Hay T.**, **Gros D.**, **Shamsi Z. and Srinivasan R.** (2022). Cross-domain detection of GPT-2-generated technical text. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States: Association for Computational Linguistics, pp. 1213–1233.

Sakhovskiy A., **Izhevskaya A.**, **Pestova A.**, **Tutubalina E.**, **Malykh V.**, **Smurov I. and Artemova E.** (2021). Rusimplesenteval-2021 shared task: Evaluating sentence simplification for russian. In *Proceedings of the International Conference "Dialogue*, pp. 607–617.

Sarvazyan A. M., **González J.Á.**, **Franco-Salvador M.**, **Rangel F.**, **Chulvi B. and Rosso P.** (2023). Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains. *Procesamiento del Lenguaje Natural* **71**, 275–288.

Schuster T., **Schuster R.**, **Shah D. J. and Barzilay R.** (2020). The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics* **46**(2), 499–510.

Schwenk H., **Chaudhary V.**, **Sun S.**, **Gong H. and Guzmán F.** (2021). WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online: Association for Computational Linguistics, pp. 1351–1361.

Scialom T., **Dray P.-A.**, **Lamprier S.**, **Piwowarski B. and Staiano J.** (2020). MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, pp. 8051–8067.

Shamardina T., **Mikhailov V.**, **Chernianskii D.**, **Fenogenova A.**, **Saidov M.**, **Valeeva A.**, **Shavrina T.**, **Smurov I.**, **Tutubalina E. and Artemova E.** (2022). Findings of the RuATD shared task 2022 on artificial text detection in Russian. *arXiv preprint arXiv:* 2206.01583.

Shavrina T. and Shapovalova O. (2017). To the methodology of corpus construction for machine learning: Taiga syntax tree corpus and parser. In *Proceedings of the "Corpora*, pp. 78–84.

Shkarin S. (2023). ruTS, a library for statistics extraction from texts in Russian.

Solaiman I., **Brundage M.**, **Clark J.**, **Askell A.**, **Herbert-Voss A.**, **Wu J.**, **Radford A.**, **Krueger G.**, **Kim J. W.**, **Kreps S.**, **McCain M.**, **Newhouse A.**, **Blazakis J.**, **McGuffie K. and Wang J.** (2019). Release strategies and the social impacts of language models.

Stiff H. and Johansson F. (2022). Detecting computer-generated disinformation. *International Journal of Data Science and Analytics* **13**(4), 363–383.

Tan R., **Plummer B. and Saenko K.** (2020). Detecting cross-modal inconsistency to defend against neural fake news. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, pp. 2081–2106.

Tang Y., **Tran C.**, **Li X.**, **Chen P.-J.**, **Goyal N.**, **Chaudhary V.**, **Gu J. and Fan A.** (2020). *Multilingual Translation with Extensible Multilingual Pretraining and Finetuning*

Temnikova I., **Marinova I.**, **Gargova S.**, **Margova R. and Koychev I.** (2023). Looking for traces of textual deepfakes in Bulgarian on social media. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, Varna, Bulgaria, Bulgaria: INCOMA Ltd., Shoumen, pp. 1151–1161.

Tiedemann J. (2012). Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey: European Language Resources Association (ELRA).

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

https://doi.org/10.1017/nlp.2024.38 Published online by Cambridge University Press

Uchendu A., Cao J., Wang Q., Luo B. and Lee D. (2019). Characterizing Man-made vs. Machine-made chatbot dialogs. In *TTO*

Uchendu A., Le T. and Lee D. (2023). Attribution and obfuscation of neural text authorship: a data mining perspective. *ACM SIGKDD Explorations Newsletter* **25**(1), 1–18.

Uchendu A., Le T., Shu K. and Lee D. (2020). Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, pp. 8384–8395.

Uchendu A., Ma Z., Le T., Zhang R. and Lee D. (2021). TURINGBENCH: A benchmark environment for turing test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP. 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 2001–2016.

Uchendu A., Mikhailov V., Lee J., Venkatraman S., Shavrina T. and Artemova E. (2022). Tutorial on artificial text detection.

Wang Y., Mansurov J., Ivanov P., Su J., Shelmanov A., Tsvigun A., Afzal O. M., Mahmoud T., Puccetti G., Arnold T., et al. (2024a). M4GT-bench: Evaluation benchmark for black-box machine-generated text detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024): Association for Computational Linguistics*, arXiv preprint arXiv: 2402.11175.

Wang Y., Mansurov J., Ivanov P., su j., Shelmanov A., Tsvigun A., Mohammed Afzal O., Mahmoud T., Puccetti G., Arnold T., Whitehouse C., Aji A. F., Habash N., Gurevych I. and Nakov P. (2024b). Semeval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, Mexico City, Mexico: Association for Computational Linguistics, pp. 2041–2063.

Wang Y., Mansurov J., Ivanov P., Su J., Shelmanov A., Tsvigun A., Whitehouse C., Mohammed Afzal O., Mahmoud T., Sasaki T., Arnold T., Aji A., Habash N., Gurevych I. and Nakov P. (2024c). M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, St. Julian's, Malta: Association for Computational Linguistics, pp. 1369–1407.

Weidinger L., Mellor J., Rauh M., Griffin C., Uesato J., Huang P.-S., Cheng M., Glaese M., Balle B., Kasirzadeh A. and et al. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:* 2112.04359.

Wolf T., Debut L., Sanh V., Chaumond J., Delangue C., Moi A., Cistac P., Rault T., Louf R., Funtowicz M., Davison J., Shleifer S., von Platen P., Ma C., Jernite Y., Plu J., Xu C., Le Scao T., Gugger S., Drame M., Lhoest Q. and Rush A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online: Association for Computational Linguistics, pp. 38–45.

Xue L., Constant N., Roberts A., Kale M., Al-Rfou R., Siddhant A., Barua A. and Raffel C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online*, Association for Computational Linguistics, pp. 483–498.

Zellers R., Holtzman A., Rashkin H., Bisk Y., Farhadi A., Roesner F. and Choi Y. (2019). Defending against neural fake news. *Advances in Neural Information Processing Systems* **32**. https://proceedings.neurips.cc/paper_files/paper/2019/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html

Zmitrovich D., Abramov A., Kalmykov A., Kadulin V., Tikhonova M., Taktasheva E., Astafurov D., Baushenko M., Snegirev A., Shavrina T., Markov S. S., Mikhailov V. and Fenogenova A. (2024). A family of pretrained transformer language models for Russian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italy: ELRA and ICCL, pp. 507–524.

Appendix A. Annotation protocols

Task overview. Choose between two judgements on the given text:

- The text is written by a human;
- The text is generated by an AI system.

Detailed task description. Follow the steps below:

- Carefully read the given text;
- Think about who could write this text;

https://doi.org/10.1017/nlp.2024.38 Published online by Cambridge University Press

- If you suppose that the text is written by a human, choose the Human option;
- If you suppose that the text is generated by an AI system, choose the AI option;

Examples.

Input text: A eto moja semja moja semja moja semja (And this is my family my family my family). Choose the AI option. The text contains obvious unnatural repetitions.

Choose the AI option. The text contains obvious unnatural repetitions.

Input text: The cat has managed to keep the behaviour pattern inherent in its wild ancestors. She is almost as good at hunting as a wild cat, but at the same time she is able to peacefully coexist with a person, show him emotional attachment, tenderness, or even show playful behaviour. (Koshka sumela sohranit' model' povedenija, prisushhuju ejo dikim predkam. Ona pochti tak zhe horosho ohotitsja, kak dikaja koshka, no v to zhe vremja sposobna mirno sosushhestvovat' s chelovekom, projavljat' k nemu jemocional'nuju privjazannost', nezhnost' ili dazhe vykazyvat' igrivoe povedenie.)

Choose the Human option. The text sounds plausible and does not contain semantic violations.

Tips. You will get texts from multiple sources and in multiple genres. Texts may look like samples from newspapers, research papers, social media. Following features help recognise generated texts:

- Inconsistent facts and incoherent writing;
- Violation of common sense and world knowledge;
- Unnecessary receptions and abrupt ending.

Following features are **NOT** helpful and can be present in human and AI texts:

- Spelling errors;
- Style fluency. Modern AI mimic human well and can write human-like texts in any genre. It is easy to be fooled by an AI system, which is able to write a research paper!

Appendix B. Human performance

Annotators show mixed results depending on the task. Generally, larger models are harder to detect than smaller ones. However, there is significant variation in accuracy when detecting different models.

https://doi.org/10.1017/nlp.2024.38 Published online by Cambridge University Press

Table B1. Human performance by models

Model	Accuracy
Human	0.832
ruGPT2-Large	0.326
ruGPT3-Small	0.674
ruGPT3-Medium	0.708
ruGPT3-Large	0.492
ruT5-Base	0.481
ruT5-Base-Multitask	0.326
ruT5-Large	0.324
mT5-Large	0.409
mT5-Small	0.789
M-BART	0.360
M-BART50	0.730
OPUS-MT	0.551
M2M-100	0.472

Cite this article: Shamardina T, Saidov M, Fenogenova A, Tumanov A, Zemlyakova A, Lebedeva A, Gryaznova E, Shavrina T, Mikhailov V and Artemova E. CoAT: Corpus of artificial texts. *Natural Language Processing* https://doi.org/10.1017/nlp.2024.38