Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5683–5691
Marseille, 20-25 June 2022

RuPAWS: A Russian Adversarial Dataset for Paraphrase Identification

Nikita Martynov*[†], Irina Krotova*[†], Varvara Logacheva[‡], Alexander Panchenko[‡], Olga Kozlova[†] and Nikita Semenov[†]
[†]MTS AI, Moscow, Russia
[‡]Skolkovo Institute of Science and Technology, Moscow, Russia
{nimarty1,ivkroto2,oskozlo9,nikita.semenov}@mts.ru, {v.logacheva,a.panchenko}@skoltech.ru

Abstract

Paraphrase identification task can be easily challenged by changing word order, e.g. as in “Can a **good** person become **bad**?”. While for English this problem was tackled by the PAWS dataset (Zhang et al., 2019), datasets for Russian paraphrase detection lack non-paraphrase examples with high lexical overlap. We present RuPAWS, the first adversarial dataset for Russian paraphrase identification. Our dataset consists of examples from PAWS translated to the Russian language and manually annotated by native speakers. We compare it to the largest available dataset for Russian ParaPhraser and show that the best available paraphrase identifiers for the Russian language fail on the RuPAWS dataset. At the same time, the state-of-the-art paraphrasing model RuBERT trained on both RuPAWS and ParaPhraser obtains high performance on the RuPAWS dataset while maintaining its accuracy on the ParaPhraser benchmark. We also show that RuPAWS can measure the sensitivity of models to word order and syntax structure since simple baselines fail even when given RuPAWS training samples.

Keywords: paraphrase detection, Russian language, dataset of paraphrases, paraphrasing

1. Introduction

Paraphrase identification task can be easily challenged by negative examples with high lexical overlap, e.g. *Can a **good** person become **bad**?* and *Can a **bad** person become **good**?* Most of the existing paraphrase identification datasets lack challenging sentence pairs that have a high bag-of-words overlap but are not paraphrases. Negative examples of this type show the significance of syntax structure and word order. Zhang et al. (2019) emphasize the importance of such examples and introduce **PAWS**, dataset constructed from Quora and Wikipedia, which consists of adversarial non-paraphrase pairs with high word overlap. Yang et al. (2019) underline the lack of such adversarial examples in existing multilingual datasets for paraphrase identification, e.g. Multi30k (Elliott et al., 2016) and Opusparcus (Creutz, 2018), and create **Cross-lingual PAWS (PAWS-X)** (Yang et al., 2019).

The same challenges exist in the paraphrasing datasets for Russian. The only existing benchmark for Russian sentential paraphrase detection **ParaPhraser** (Pivovarova et al., 2017) contains sentence pairs from news headlines and also lacks such challenging examples (See Figure 1). Table 1 shows, that the models trained on ParaPhraser tend to classify the non-paraphrase examples with high BOW overlap as a paraphrase. To bridge this gap and boost the research of Russian paraphrase identification, we create the **Russian PAWS (RuPAWS)** dataset, the first Russian adversarial dataset for paraphrase classification with high lexical overlap. Our new corpus consists of 8,814 examples from PAWS translated to Russian, cleaned and manually annotated by native speakers.

We show that state-of-the-art model for paraphrase identification RuBERT (Kuratov and Arkhipov, 2019) trained on ParaPhraser fails on RuPAWS benchmark. By contrast, when adding RuPAWS examples RuBERT substantially improves its performance on challenging sentence pairs without significantly reducing performance on ParaPhraser.

We also provide the evaluation of baseline models and show that RuPAWS can measure the sensitivity of models to word order and syntax structure.

The RuPAWS dataset, including both test and train sets, and RuBERT trained on both ParaPhraser and RuPAWS will be released publicly at github[1].

Our contributions are as follows:

- We present RuPAWS - the first open adversarial paraphrase identification dataset for Russian, with high number of negative examples with high lexical overlap.
- We conduct an evaluation of baseline and state-of-the-art models and demonstrate that RuPAWS can measure the sensitivity of models to word order and syntax structure of Russian language.
- We show that adding RuPAWS training data can substantially improve the performance of state-of-the-art models and make them more robust to real-world examples without significantly reducing their performance on previous benchmarks.

2. Related work

According to Androutsopoulos and Malakasiotis (2010) computational methods of paraphrasing are presented in three main task forms: paraphrase identification (or

* equal contribution

[1] https://github.com/mts-ai/rupaws-dataset

Sentence 1	**Sentence 2**	**Type**	**RuBERT (PP)**	**RuBERT (PP+RuPAWS)**
Можно ли **хорошему** человеку стать **плохим**? *Can a* ***good*** *person become* ***bad****?*	Можно ли **плохому** человеку стать **хорошим**? *Can a* ***bad*** *person become* ***good****?*	Adjective swap	0.96	0.02
У какой авиакомпании есть дешевый перелет из **Амстердама** в **Джакарту**? *Which airline has cheap flight from* ***Amsterdam*** *to* ***Jakarta****?*	У какой авиакомпании дешевые перелеты из **Джакарты** в **Амстердам**? *What airline has cheap flight from* ***Jakarta*** *to* ***Amsterdam****?*	Named entity swap	0.97	0.08
Очередное исполнение оперы Карла Оге Расмуссена было **записано** в 2005 году и **опубликовано** в 2006 году. *A further completion of the opera, by Karl Aage Rasmussen, was* ***recorded*** *in 2005 and* ***published*** *in 2006 .*	Еще одна экранизация оперы Карла Оге Расмуссена была **опубликована** в 2005 году и **записана** в 2006 году. *Another completion of the opera, by Karl Aage Rasmussen, was* ***published*** *in 2005 and* ***recorded*** *in 2006.*	Verb swap	0.96	0.03
Эвари Байзо (3 июня 1821 - 6 февраля 1910 - Нант) - французский военный **физиолог**. *Evariste Baizeau (June 3, 1821 - February 6, 1910, Nantes) was a French military* ***physician****.*	Эвари Байзо (3 июня 1821 - 6 февраля 1910 - Нант) - французский военный **физик**. *Evariste Baizeau (June 3, 1821 - February 6, 1910, Nantes) was a French military* ***physicist****.*	Word replacement	0.96	0.02

Table 1: Examples of non-paraphrases with high lexical overlap and corresponding scores by RuBERT trained on ParaPhraser (**PP**) and RuBERT trained on ParaPhraser + RuPAWS (**PP + RuPAWS**). Scores represent estimate of probability that these two sentences are paraphrases. All examples are from the test part of RuPAWS. **Type** column shows the type of adversarial non-paraphrase generation, boldface indicates the differences between two sentences.

detection), paraphrase generation, and paraphrase extraction (e.g. from corpora). Paraphrase identification (Socher et al., 2011; Zhang and Patrick, 2005; Jia et al., 2020) finds applications in many branches of natural language processing, such as machine translation (Callison-Burch et al., 2006; Madnani et al., 2007; Mayhew et al., 2020; Apidianaki et al., 2018), plagiarism detection (Hunt et al., 2019), text summarization (Mani and Maybury, 2001; Zhang et al., 2017) including sense compression (Napoles et al., 2011). Question paraphrasing is a necessary part of knowledge-based question answering (QA) systems (Fader et al., 2014; Yin et al., 2015), which can be used to retrieve relevant documents and passages or when mapping user questions to list of a frequently asked questions (Tomuro, 2003). The paraphrase identification task is closely linked with paraphrase generation in the field of language resources, as many datasets have been created for training and evaluating both types of models.

Sentential paraphrase datasets for English. Most of the work on paraphrase identification focuses primarly on English. This task inspired a number of SemEval evaluation tasks in 2012 (Agirre et al., 2012), 2013 (Hendrickx et al., 2013), 2015 (Xu et al., 2015) and 2016 (Agirre et al., 2016), which have encouraged the development of baseline decisions and their further improvement.

There also exist various paraphrase datasets for English, which consist of human-labeled sentential paraphrases as the dataset we release (See Table 2 for detailed comparison).

Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005) contains 5,801 sentence pairs, each human labeled with a binary judgment as to whether the pair is a paraphrase. Sentence-level paraphrases were selected from a large corpus of topic-clustered news data through the use of heuristic extraction techniques in conjunction with an SVM-based classifier. Next, the collected pairs were submitted to human annotators, who judged 67% of the original pairs as semantically equivalent. *The Quora Question Pairs (QQP)* includes 404290 question pairs from the Quora website, each annotated with a binary value indicating whether the two questions are a paraphrase of each other. MRPC and QQP are part of the GLUE similarity and paraphrase task (Wang et al., 2018). *Twitter News URL Corpus (TURL)* (Lan et al., 2017) consists of 51,524 manually labeled sentence pairs captured from Twitter by linking tweets through shared URLs. The captured sentence pairs were given a similarity score ranging from 1 to 6.

Paraphrasing datasets for Russian. Russian is less represented in paraphrase research concerning resource development and algorithm evaluation. Two evaluation tasks focused on paraphrasing: AINL 2016 Paraphrase Detection Shared Task based on *ParaPhraser* corpus (Pivovarova et al., 2017) and Dialogue Paraphrased Plagiarism Detection Competition in 2017 based on *ParaPlag* corpus with main focus on text-level rephrasing (Sochenkov et al., 2017). In addition, Gudkov et al. (2020) presented the ParaPhraser Plus corpus distilled from a database of news headlines. There is also a part of some multilingual paraphrase resources, such as *Opuscarpus* (Creutz, 2018) and *PPDB* (Ganitkevitch et al., 2013). The Paraphrase Detection Shared Task attracted attention to the Russian paraphrase identification and led to further research. Kuratov and Arkhipov (2019) show that finetuning a monolingual BERT-based model (RuBERT) on the ParaPhraser corpus yields better results than previous approaches. In our work, we show that the performance of this model on complex sentence pairs with high-lexical overlap can be improved by adding our RuPAWS corpus to the training set.

Adversarial paraphrasing datasets. There is also a limited number of datasets with adversarial paraphrase examples, whose objective is to highlight the deficiencies of state-of-the-art models. Our work is based on the

	MRPC	QQP	TURL	PAWS	PAWS-X	ParaPhraser	RuPAWS (ours)
Language	English	English	English	English	Multilingual	**Russian**	**Russian**
Size (sentence pairs)	5 801	404 290	51 524	108 463	320 065	9 151	8 814
% Positive class	18	37	25	33	44	63	39
Type	News	Social	Social	**Social + Wiki**	**Social + Wiki**	News	**Social + Wiki**
Adversarial examples	No	No	No	**Yes**	**Yes**	No	**Yes**
Manual annotation	**Yes**	**Yes**	**Yes**	**Yes**	Dev&Test	**Yes**	**Yes**

Table 2: Comparison of the existing sentential datasets for paraphrase detection.

	PAWS$_{QQP}$	**PAWS**$_{Wiki}$
# **Raw pairs**	12 663	95 798
Noise-filtered pairs		
Total # pairs	12 225	87 695
paraphrase	4 157	31 570
non-paraphrase	8 068	56 125
Machine-translated pairs		
Total # pairs	6 076	38 558
paraphrase	2 082	13 967
non-paraphrase	3 994	24 591
Annotated pairs		
Total # pairs	2 154	6 660
paraphrase	9 07	2 563
non-paraphrase	1 247	4 097

Table 3: The number of sentence pairs on all stages of RuPAWS creation.

PAWS dataset (Zhang et al., 2019), which contains paraphrase and non-paraphrase pairs with high-lexical overlap. (Yang et al., 2019) present the cross-lingual dataset *PAWS-X*, an extension of the PAWS examples to six languages: Spanish, French, German, Chinese, Japanese, and Korean. Conversely, Nighojkar and Licato (2021) introduce the *Adversarial Paraphrasing Task*, which goal is to provide semantically equivalent but lexically and syntactically desperate paraphrases.

3. Dataset

The RuPAWS dataset creation is based on the machine translation of the original PAWS corpus from English to Russian and the further annotation of the resulting sentence pairs. Following Yang et al. (2019), we choose translation instead of repeating the PAWS data generation approach. Due to human resource constraints, the dataset was machine translated and then reviewed by human annotators, who are Russian native speakers. We also perform the data cleaning procedure before and after the machine translation stage to reduce the number of noisy sentences. As a result, we select 8814 human annotated translations of paraphrase and non-paraphrase pairs. For the detailed statistics on the amount of data on each stage please see Table 3.

3.1. Data denoising

The original PAWS dataset was automatically generated by two methods: word swapping and back-translation. Consequently, it contains a number of noisy or incoherent sentences which can yield ungrammatical sentences, for example *I just turned 13 and I was I 5* or *As , he is the and can borrow the Inromaru to become*.

To eliminate the sentences of poor quality, we first remove the sentences with non-ASCII characters. Next, we provide a perplexity-based selection (Lin et al., 1997; Gao et al., 2002) for the remaining sentence pairs by calculating the perplexity of each sentence with a large transformer-based GPT-2 model (Radford et al., 2019). We select the sentence pairs with a perplexity score lower than 7. In addition, PAWS contants some sentences which are not semantically similar. We use a multilingual BERT embedding model LaBSE (Feng et al., 2020), which establishes a new state of the art on multiple parallel text retrieval tasks, and rank the sentence pairs by a cosine similarity between LaBSE representations of the sentences. We remove the sentence pairs with a similarity score lower than 0.9.

3.2. Machine translation

We use facebook/wmt19-en-ru model from Ng et al. (2019), the Facebook FAIR's submission to the WMT19 shared news translation task, for machine translation from English to Russian. However, different translation strategies work better for different sentences, so we use greedy inference, beam search, and top-k sampling to form a pool of possible sentence translations. Finally, we select the best candidate by cosine similarity between LaBSE embeddings of the original and translated sentences. In addition, we filter the resulting Russian sentence pairs with respect to perplexity of sentences and cosine similarity scores between their LaBSE representations, as described in Section 3.1.

3.3. Human annotation

Even after the automatic post-processing, the machine-translated data may contain noise and errors. We ask in-house annotators to evaluate the resulting sentence pairs for meaning preservation and the correctness of the text. In order to make the best use of limited resources, we selected the best sentence pairs and passed them to the annotators. It takes three 40-hours working weeks to annotate 10,119 sentence pairs by two raters. The sentence pairs were judged under the following criteria:

1. Both sentences are coherent, grammatically, and lexically correct, they are readable and their meaning is well understood.

Model	PP		RuPAWS					
			Wiki		QQP		Wiki+QQP	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
BOW								
PP	62.5	60.5	34.7	54.9	45.4	60.2	35.0	55.5
PP + RuPAWS*	62.5	60.0	46.0 (+11.3)	55.6 (+0.7)	49.2 (+3.8)	61.0 (+0.8)	46.3 (+11.3)	56.0 (+0.5)
BiLSTM								
PP	74.3	81.6	39.1	54.4	46.0	62.0	40.1	55.6
PP + RuPAWS*	66.3	74.1	65.4 (+26.3)	60.9 (+6.5)	52.4 (+6.4)	60.0	63.5 (+23.4)	60.7 (+5.1)
RuBERT								
PP	**85.0**	**87.7**	38.4	55.5	46.0	63.0	39.1	56.2
PP + RuPAWS*	**84.6**	**87.3**	**79.6** (+41.2)	**75.4** (+19.2)	**73.6** (+27.6)	**71.3** (+8.3)	**79.0** (+39.9)	**74.9** (+18.7)

Table 4: **Acc**uracy (%) of classification and **F**$_1$ (%) scores on ParaPhraser (**PP**) and training sets. PP indicates that the model is trained on the ParaPhraser training set and PP + RuPAWS shows that the model is trained on both ParaPhraser and RuPAWS training sets. Column names indicate the test set. **Wiki**pedia and Quora Question Pairs (**QQP**) stand for the RuPAWS test set divided into separate parts, **Wiki+QQP** stands for the concatenated test set. Numbers in parenthesis indicate gains from adding RuPAWS to training data. * - RuPAWS as training data stands for the training sets from both Wiki and QQP parts of RuPAWS.

2. Whether or not the sentences are paraphrases of each other.

Even though the original sentence pairs from PAWS are labeled as paraphrases and non-paraphrases, these labels may become incorrect after the machine-translations stage of the dataset creation. Therefore we annotate whether the sentence pairs are paraphrases or not. We accept only correct sentence pairs with full annotator agreement on both criteria (87% of all annotated data).

3.4. Resulting dataset

We created RuPAWS – the first Russian dataset for paraphrase detection with a high number of adversarial non-paraphrase pairs with high word overlap. Our dataset contains 8,814 manually annotated sentence pairs from both parts of the PAWS dataset (QQP and Wiki). Non-paraphrases account for 61% of the total amount of RuPAWS. 47% of RuPAWS non-paraphrases have n-gram overlap over 0.5 in contrast to the ParaPhraser, where only 6% of non-paraphrases have overlap over 0.5 (See Figure 1).

4. Evaluation

To evaluate the RuPAWS dataset, we use it to train the paraphrase identification models.

4.1. Models

The goal of RuPAWS is to investigate the models' ability to capture the sentence structure and word order. As discussed previously, paraphrase identification models tend to classify the sentence pair with high lexical overlap as a paraphrase. In our work, we consider three different models with varying complexity and expressiveness: two baseline encoders and one advanced model, that achieved state-of-the-art performance on Russian paraphrase identification.

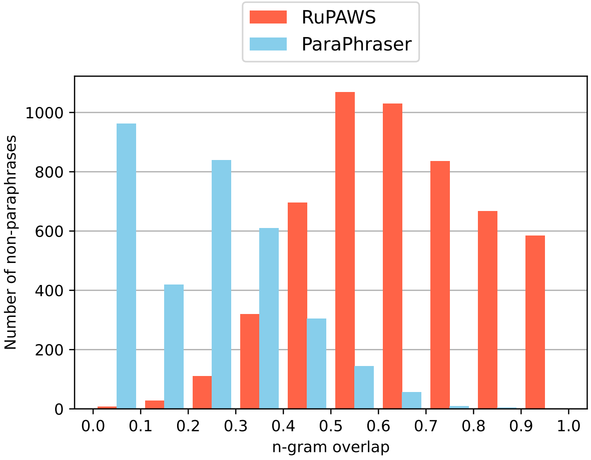

Figure 1: Distribution of non-paraphrases over n-gram overlap.

The first baseline is a bag-of-words (**BOW**) based on token unigram and bigram encoding. The second one is a bi-directional LSTM (**BiLSTM**) that produces a contextualized sentence encoding. In contrast to the simple BOW model, BiLSTM captures non-local contexts. For BiLSTM, we use pre-trained FastText word embeddings from Mikolov et al. (2018) and keep them frozen during training. We used bi-directional LSTM with hidden size 64 and calculate sentence embedding as the last hidden state. For both BOW and BiLSTM, we calculate cosine similarity between sentence embeddings and treat value above 0.5 as a paraphrase. Finally, we evaluate **RuBERT** (Kuratov and Arkhipov, 2019), a deep bidirectional pre-trained monolingual transformer, that obtained state-of-the-art results on paraphrase identification task in Russian. Similar to (Kuratov and Arkhipov, 2019), when fine-tuning RuBERT, we encode both sentences jointly and classify embedding of the [CLS] token.

4.2. Experiments and Results

Our goal is to understand how well the models trained on ParaPhraser generalize to RuPAWS challenging pairs and to test how well the selected models are able to learn on RuPAWS. We also test how well models trained on both ParaPhraser and RuPAWS perform on ParaPhraser. We use public implementation of RuBERT[2] model trained on the Russian part of Wikipedia and news data and fine-tune RuBERT with learning rate $2e^{-5}$.

Table 4 shows results on the RuPAWS and ParaPhraser benchmarks. The models are trained on the ParaPhraser training examples and on the combination of ParaPhraser and RuPAWS. When training on two datasets, we sample instances from them in random order. The columns show the results of different test sets.

BOW classifier is the simplest baseline model and considers only local context information, so it has the worst performance and shows almost no improvement from new examples when trained on ParaPhraser and RuPAWS jointly.

BiLSTM outperforms BOW in almost all cases, but its performance is still lower than RuBERT scores. State-of-the-art model **RuBERT** trained on ParaPhraser and RuPAWS shows the highest performance scores and substantial gains when added RuPAWS examples without significantly reducing its performance on the ParaPhraser test set. Therefore, performance changes on RuPAWS are more visible if the models are trained on both datasets. There is no significant difference between BiLSTM and RuBERT trained on ParaPhraser when testing on RuPAWS challenging pairs. In contrast, adding RuPAWS samples to the training set lead to the gain of up to 41%. For example, the difference between BiLSTM and RuBERT trained on ParaPhraser is 0.6% on RuPAWS test, but it rises to 14.2% when trained on ParaPhraser and RuPAWS. There is also almost no negative impact on performance on ParaPhraser. Therefore, RuPAWS trained on ParaPhraser and RuPAWS shows the best performance on both datasets.

5. Error analysis

Despite the fact that the RuBERT model trained on both RuPAWS and ParaPhraser datasets shows the best performance, it still fails to reach F_1-score higher than 0.8. We perform a broad error analysis to find out the remaining problems and possible ways to improve our data and the model performance. Considering the different nature of paraphrase and non-paraphrase sentence pairs we inspect false negative and false positive classifications. We analyze common paraphrase and non-paraphrase types represented in our test data and their impact on the error distributions. Moreover, we add additional observations about the impact of named entities translation we encountered during our manual analysis.

Our test set consists of 741 paraphrase and 932 non-paraphrase pairs. The model results for 218 false positives and 130 false negatives. Both classes are of particular importance for paraphrase identification on real-world tasks. We perform our analysis on all of these samples and manually annotate 1,673 sentence pairs.

5.1. False positive errors

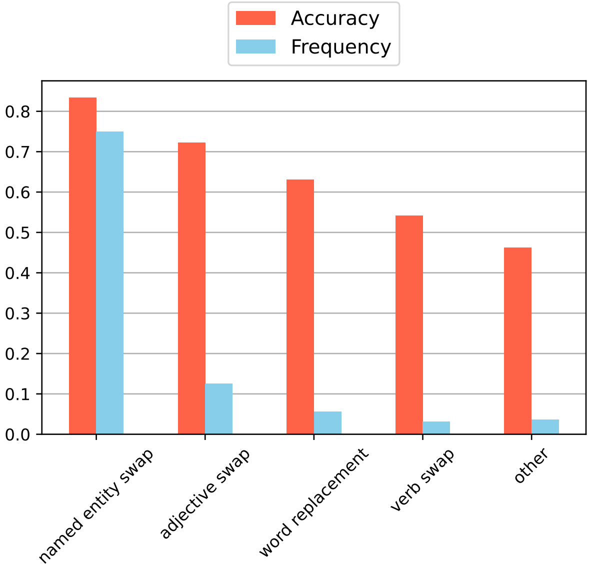

Figure 2: Accuracy scores and frequency for non-paraphrase types presented in the RuPAWS test set.

Owning to the fact that the RuPAWS examples are PAWS samples translated into Russian, our manual annotation of non-paraphrase is based on PAWS example generation strategies. The PAWS automatic generation method is based on two techniques. The first one swaps the words to generate a sentence pair with the same BOW. This method tends to produce non-paraphrase sentence pairs. The second strategy is based on back-translation and usually produces paraphrases. It is not possible to completely transfer these sentence pair classes to our test data since additional lexical and syntax changes arise due to the following machine translation to Russian. However, we use the description of these techniques as a basis for our error analysis.

Our annotation scheme includes the following categories: **named entity swap**, **verb swap**, **adjective swap**, **word replacement** and **other**, which includes infrequent non-paraphrase types, such as noun phrase swap, word deletion, if it has a significant influence on the sentence meaning, or combination of different sentence changes. The examples of these non-paraphrase categories are presented in Table 1.

In order to estimate the distribution of classes in the non-paraphrase part of the test set and its relationship to false positive errors we manually annotate the non-paraphrase sentences in the test set with the type of change operation.

[2] http://docs.deeppavlov.ai/en/master/features/models/bert.html

The obtained accuracy scores and class frequencies for each category are shown in Figure 2. The prevailing non-paraphrase generation type is the **named entity swap** (699 pairs, 75% of all non-paraphrases in the test set) and the RuBERT model trained on both datasets achieves the best accuracy (83%) on this class. The second most frequent class is the **adjective swap** (121 pairs, 13%), and the model achieves the second best accuracy score (72%) on it. The remaining three classes account for 12% of non-paraphrase pairs in the test set, while the accuracy score for them is below 65%.
We attribute these scores to an unequal distribution of non-paraphrase generation types and the domain of the original PAWS dataset, which contributes to the prevalence of the sentence pairs with named entities. We assume, that the addition of underrepresented non-paraphrase pairs will improve the quality of classification.

5.2. False negative errors

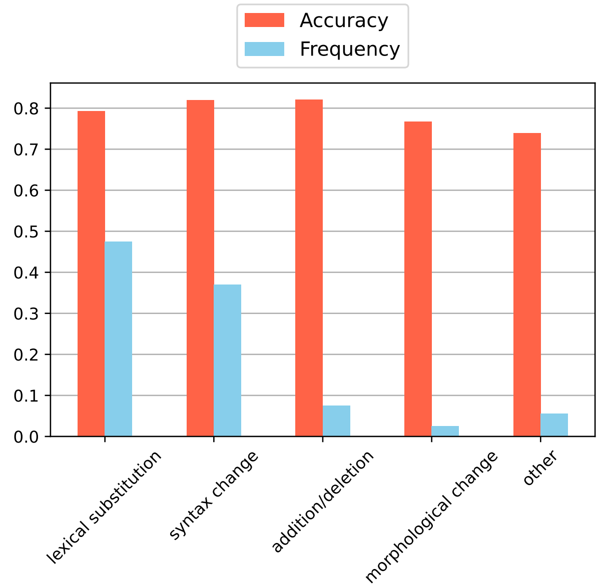

Figure 3: Accuracy scores and frequency for paraphrase types presented in the RuPAWS test set.

Different paraphrasing strategies In our analysis of false negative errors, we rely on the paraphrase typology proposed by Vila et al. (2014). According to them, we divide the paraphrase sentence pairs from out test set into five categories:

- **lexical substitution**: Он умер в 1909 году в Монреале и был **погребен** в Калгари. *He died in 1909 in Montreal and was* ***buried*** *in Calgary.* / Он умер в 1909 году в Монреале и был **похоронен** в Калгари. *He died in 1909 in Montreal and was* ***entombed*** *in Calgary.*

- **syntax-based changes**: Aker Yards приобрела STX Europe **в 2008 году.** *STX Europe was acquired by Aker Yards* ***in 2008.*** / **В 2008 году** Aker Yards приобрела компания STX Europe. ***In 2008*** *Aker Yards was acquired by STX Europe.*

- **morphological changes**: Фильм срежиссировал Джон Тейлор, а **спродюсировал** Стюарт Легг. *The film was directed by John Taylor and* ***produced*** *by Stuart Legg.* / Фильм срежиссировал Джон Тэйлор, а **продюсером** стал Стюарт Легг. *The film was directed by John Taylor and Stuart Legg was* ***a producer.***

- **addition/deletion**: "Потез X" - французский транспортный самолет **общего назначения** 1920-х годов, спроектированный и построенный фирмой Potez. *The Potez X was a French 1920s* ***general-purpose*** *colonial transport aircraft designed and built by Potez.*/ "Потез X" - французский транспортный самолет 1920-х годов, спроектированный и построенный компанией Potez. *The Potez X was a French 1920s colonial transport aircraft designed and built by Potez.*

- **other**, which includes abbreviated words, change of word format or combination of different paraphrasing strategies: Саммерби родился **в г. Сиренчестер в Англии** и умер **в г. Винчестер в Англии**. *Summerbee was born* ***in Cirencester in England*** *and died* ***in Winchester in England.*** / Саммерби родился **в городе Сиренчестер, Англия**, и умер **в Винчестере, Англия**. *Summerbee was born* ***in the Cirencester city, England*** *and died* ***in Winchester, England.***

We manually annotate 741 paraphrase pairs in the test set for the above-mentioned paraphrase categories. The obtained accuracy scores and class frequencies are shown in Figure 3. The most frequent paraphrase generation classes are lexical substitution (47%) and syntax change (37%), but the relationship between the category frequency and its accuracy score is not obvious. The addition/deletion paraphrasing type has the highest accuracy score (82%), and the lowest is 74% for the "other" category.
Named entities translation However, we extend the list of possible error classes with an additional class which we encountered during our manual annotation. The RuPAWS dataset was obtained by a machine translation from English to Russian, which has raised additional difficulties for the paraphrase classification models. As mentioned in Section 5.1 the dataset contains a large number of sentences with named entities. Russian and English languages have different graphic systems, thus when translating named entities different techniques can be used: translation, transliteration, or transcription. For example, the football club *"Red Star Belgrade"* can be translated as *белградская "Red star"*, *"Red Star Belgrade"*, *"Red Star" из Белграда*, *"Красная звезда" Белграда*. As another example, *David Burtka* as *Дэвид*

Бартка, *Давид Буртка* or *Дэвид Бертка*. We manually annotate false negatives and find different named entity translations in 52% of paraphrase pairs. We assume, that an option for improving the paraphrase classification model is the further work on this particular problem.

6. Conclusion

We introduce RuPAWS, the first adversarial paraphrase identification dataset for Russian with 8,814 human annotated sentence pairs with high lexical overlap. We compare our dataset to the ParaPhraser, the largest available dataset for Russian, and show that the best available state-of-the-art model for paraphrase identification RuBERT trained on ParaPhraser fails to solve many examples of the RuPAWS dataset. By contrast, when adding RuPAWS examples RuBERT improves its accuracy to 79% while maintaining performance on the ParaPhraser benchmark. We also conduct an evaluation of paraphrase identifiers and demonstrate that RuPAWS can measure the sensitivity of models to word order and syntax structure of Russian language.

7. Bibliographical References

Agirre, E., Cer, D., Diab, M., and Gonzalez-Agirre, A. (2012). Semeval-2012 task 6: A pilot on semantic textual similarity. In * *SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393.

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez Agirre, A., Mihalcea, R., Rigau Claramunt, G., and Wiebe, J. (2016). Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511.* ACL (Association for Computational Linguistics).

Androutsopoulos, I. and Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

Apidianaki, M., Wisniewski, G., Cocos, A., and Callison-Burch, C. (2018). Automated paraphrase lattice creation for hyter machine translation evaluation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 480–485.

Callison-Burch, C., Koehn, P., and Osborne, M. (2006). Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24.

Creutz, M. (2018). Open subtitles paraphrase corpus for six languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Dolan, W. B. and Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016). Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*.

Fader, A., Zettlemoyer, L., and Etzioni, O. (2014). Open question answering over curated and extracted knowledge bases. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 1156–1165, New York, NY, USA. Association for Computing Machinery.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2020). Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013). Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764.

Gao, J., Goodman, J., Li, M., and Lee, K.-F. (2002). Toward a unified approach to statistical language modeling for chinese. *ACM Transactions on Asian Language Information Processing (TALIP)*, 1(1):3–33.

Gudkov, V., Mitrofanova, O., and Filippskikh, E. (2020). Automatically ranked Russian paraphrase corpus for text generation. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 54–59, Online, July. Association for Computational Linguistics.

Hendrickx, I., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Szpakowicz, S., and Veale, T. (2013). SemEval-2013 task 4: Free paraphrases of noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 138–143, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Hunt, E., Janamsetty, R., Kinares, C., Koh, C., Sanchez, A., Zhan, F., Ozdemir, M., Waseem, S., Yolcu, O., Dahal, B., Zhan, J., Gewali, L., and Oh, P. (2019). Machine learning models for paraphrase identification and its applications on plagiarism detection. In *2019 IEEE International Conference on Big Knowledge (ICBK)*, pages 97–104.

Jia, X., Zhou, W., Sun, X., and Wu, Y. (2020). How to ask good questions? try to leverage paraphrases. In *Proceedings of the 58th Annual Meeting of the Asso-*

ciation for Computational Linguistics, pages 6130–6140.

Kuratov, Y. and Arkhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.

Lan, W., Qiu, S., He, H., and Xu, W. (2017). A continuously growing dataset of sentential paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1224–1234, Copenhagen, Denmark, September. Association for Computational Linguistics.

Lin, S.-C., Tsai, C.-L., Chien, L.-F., Chen, K.-J., and Lee, L.-S. (1997). Chinese language model adaptation based on document classification and multiple domain-specific language models. In *Fifth European Conference on Speech Communication and Technology*.

Madnani, N., Ayan, N. F., Resnik, P., and Dorr, B. (2007). Using paraphrases for parameter tuning in statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 120–127.

Mani, I. and Maybury, M. T. (2001). Automatic summarization.

Mayhew, S., Bicknell, K., Brust, C., McDowell, B., Monroe, W., and Settles, B. (2020). Simultaneous translation and paraphrase for language education. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 232–243.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Napoles, C., Callison-Burch, C., Ganitkevitch, J., and Van Durme, B. (2011). Paraphrastic sentence compression with a character-based metric: Tightening without deletion. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 84–90.

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., and Edunov, S. (2019). Facebook fair's wmt19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319.

Nighojkar, A. and Licato, J. (2021). Improving paraphrase detection with the adversarial paraphrasing task. *arXiv preprint arXiv:2106.07691*.

Pivovarova, L., Pronoza, E., Yagunova, E., and Pronoza, A. (2017). Paraphraser: Russian paraphrase corpus and shared task. In *Conference on Artificial Intelligence and Natural Language*, pages 211–225. Springer.

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

Sochenkov, I., Zubarev, D., and Smirnov, I. (2017). The paraplag: Russian dataset for paraphrased plagiarism detection. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue*, volume 1, pages 284–297.

Socher, R., Huang, E. H., Pennington, J., Ng, A. Y., and Manning, C. D. (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *NIPS*, volume 24, pages 801–809.

Tomuro, N. (2003). Interrogative reformulation patterns and acquisition of question paraphrases. In *Proceedings of the second international workshop on Paraphrasing*, pages 33–40.

Vila, M., Martí, M. A., Rodríguez, H., et al. (2014). Is this a paraphrase? what kind? paraphrase boundaries and typology. *Open Journal of Modern Linguistics*, 4(01):205.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November. Association for Computational Linguistics.

Xu, W., Callison-Burch, C., and Dolan, W. B. (2015). Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit). In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 1–11.

Yang, Y., Zhang, Y., Tar, C., and Baldridge, J. (2019). PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China, November. Association for Computational Linguistics.

Yin, P., Duan, N., Kao, B., Bao, J., and Zhou, M. (2015). Answering questions with complex semantic constraints on open knowledge bases. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1301–1310.

Zhang, Y. and Patrick, J. (2005). Paraphrase identification by text canonicalization. In *Proceedings of the Australasian Language Technology Workshop 2005*, pages 160–166.

Zhang, C., Sah, S., Nguyen, T., Peri, D., Loui, A., Salvaggio, C., and Ptucha, R. (2017). Semantic sentence embeddings for paraphrasing and text summarization. In *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 705–709. IEEE.

Zhang, Y., Baldridge, J., and He, L. (2019). PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* *Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota, June. Association for Computational Linguistics.

8. Language Resource References