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

Main Contributions

- The first open **adversarial paraphrase identification dataset for Russian**, with high number of negative examples with high lexical overlap;
- An **evaluation of baseline and state-of-the-art models**, which demonstrate that RuPAWS can measure the sensitivity of models to word order and syntax structure of Russian language;
- 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**.

Comparison of the existing sentential datasets for paraphrase detection

	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**

English translation of 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)

Sentence 1	Sentence 2	Type	RuBERT (PP)	RuBERT (PP+RuPAWS)
*Can a **good** person become **bad**?*	*Can a **bad** person become **good**?*	Adjective swap	0.96	0.02
*Which airline has cheap flight from **Amsterdam** to **Jakarta**?*	*What airline has cheap flight from **Jakarta** to **Amsterdam**?*	Named entity swap	0.97	0.08
*A further completion of the opera, by Karl Aage Rasmussen, was **recorded** in 2005 and **published** in 2006 .*	*Another completion of the opera, by Karl Aage Rasmussen, was **published** in 2005 and **recorded** in 2006.*	Verb swap	0.96	0.03
*Evariste Baizeau (June 3, 1821 - February 6, 1910, Nantes) was a French military **physician**.*	*Evariste Baizeau (June 3, 1821 - February 6, 1910, Nantes) was a French military **physicist**.*	Word replacement	0.96	0.02

RuPAWS Creation

- 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.
- Due to human resource constraints, the dataset was machine translated by **facebook/wmt19-en-ru model** 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.

RuPAWS Creation

	$PAWS_{QQP}$	$PAWS_{Wiki}$
# 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

Evaluation

Model	PP		Wiki		RuPAWS 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	55.6	49.2	61.0	46.3	56.0
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	60.9	52.4	60.0	63.5	60.7
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**	**75.4**	**73.6**	**71.3**	**79.0**	**74.9**

Models

- 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. We use **pre-trained FastText** word embeddings 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,** a deep bidirectional pre-trained monolingual transformer, that obtained **state-of-the-art** results on paraphrase identification task in Russian. We encode both sentences jointly and classify embedding of the [CLS] token.

Links & Contacts

- GitHub: `https://github.com/mts-ai/rupaws-dataset`
- For all questions contact: `nimarty1@mts.ru`, `i.krotova@mts.ai`

Skoltech
Skolkovo Institute of Science and Technology
MTS AI