Записки научных
семинаров ПОМИ
Том 529, 2023 г.

D. Karpov, M. Burtsev

MONOLINGUAL AND CROSS-LINGUAL KNOWLEDGE TRANSFER FOR TOPIC CLASSIFICATION

Abstract. In this work, we investigate knowledge transfer from the `RuQTopics` dataset. This Russian topical dataset combines a large number of data points (361,560 single-label, 170,930 multi-label) with extensive class coverage (76 classes). We have prepared this dataset from the "Yandex Que" raw data. By evaluating the models trained on `RuQTopics` on the six matching classes from the Russian `MASSIVE` subset, we show that the `RuQTopics` dataset is suitable for real-world conversational tasks, as Russian-only models trained on this dataset consistently yield an accuracy around 85% on this subset. We have also found that for the multilingual BERT trained on `RuQTopics` and evaluated on the same six classes of `MASSIVE` (for all `MASSIVE` languages), the language-wise accuracy closely correlates (Spearman correlation 0.773 with p-value 2.997e-11) with the approximate size of BERT pretraining data for the corresponding language. At the same time, the correlation of language-wise accuracy with the linguistic distance from the Russian language is not statistically significant.

§1. Introduction

As the natural language processing (NLP) field continues to progress, applications of chatbots and virtual assistants are becoming increasingly popular and widespread. These applications can assist with a wide range of tasks, from answering simple questions to making appointments and providing emotional feedback [33].

Building a virtual assistant is not a trivial task. A typical dialogue system has complex configuration and consists of four main components. The Natural Language Understanding component maps natural language utterances to a labeled semantic representation. The Dialogue Manager keeps track of the dialogue state and maintains the conversation flow. The Natural Language Generation component translates semantic representation into natural language utterances. The Natural Language Understanding

Key words and phrases: dataset and topic classification and knowledge transfer and cross-lingual knowledge transfer.

component joins a variety of NLP models including classification of the sentiment, topics, and intents of the user's utterances [19] into the dialogue system.

Collecting and labeling conversational datasets requires tremendous effort [16]. To the best of our knowledge, existing work lacks conversational topical datasets for the Russian language. Moreover, existing Russian topical datasets suffer from various problems: some of them cover a hopelessly insufficient number of topics, some datasets lack samples, while others are either too specific or lack conversational samples. Additionally, knowledge transfer for topical datasets is especially under-researched, even though it can be particularly helpful for lower-resource languages [17].

In this study, we explore the Russian topical dataset `RuQTopics`, which consists of questions and summarized answers of the users from "Yandex Que", a Russian question-answering website. Every question belongs to one or several of the 76 "Yandex Que" topics. We have carefully selected these topics by looking at the DREAM dialog system requirements [2, 21]. We prove that this dataset is suitable for conversational tasks. This dataset has a single-label part as well as a multi-label part, and even the single-label part of `RuQTopics` by far outsizes all other Russian topical datasets that can be used for conversational topic classification. We have also studied cross-lingual knowledge transfer from our Russian dataset to 50 different languages on parallel conversational data from the `MASSIVE` dataset.

§2. Related Work

Plenty of topical datasets have been made available by the research community. However, not all of these datasets are well-suited for conversational tasks. The majority of topical datasets consist of large pieces of written text (mostly news). Training on these datasets makes models overfit on long pieces of data, which can lead to poor performance on conversational utterances. Moreover, the class nomenclature in these datasets is usually quite small, and, therefore, a vast majority of topics one can bring up in conversation are still out of their coverage. Furthermore, these datasets rarely contain Russian utterances. As an example of such a dataset we note `AG-NEWS` [32], which has only four topics, or the dataset from `The Guardian` [29]. These datasets are also English-only.

The news dataset `MLSUM` [28] has versions for several different languages (French, German, Turkish, Spanish, Russian). Due to the large size of news articles (compared to conversational utterances), examples in this dataset

are too large for conversational tasks. Moreover, the 16 Russian topical classes from this dataset are derived from news categories. These classes still do not cover a vast majority of conversational topics.

The same problem of text length also holds for the `XGLUE-nc` [23] dataset. This 10-class news dataset has an English-only training set and a test sample from five European languages, including Russian. An ontology dataset `DBpedia` [22] also suffers from this issue as it contains very long texts. Moreover, the nomenclature of this dataset (14 classes) is by no means sufficient for topical classification.

Other topical datasets are too domain-specific, and thus they are a poor fit for general-purpose tasks. Among such datasets, we can mention `LexGLUE` [6] and `LEXTREME` [24] benchmarks, which are focused on legal-specific topics. Other datasets have been created for patent classification [30] and book title classification [5]. Russian datasets that were created for the classification of reviews on Russian medical facilities [4] or classification of university-specific intents [25] can also be included in this category. However, the majority of conversational datasets are also very domain-specific, for example, conversational `NegoChat` dataset for the negotiation domain [14].

We can also mention the product review dataset from Amazon [11]. This dataset contains reviews of products sold on Amazon from different categories, grouped by the topic. However, the topics provided in this dataset are also insufficient for building a general-purpose topic classifier, as the possible range of topics to discuss differs from the variety of Amazon product categories. Additionally, the dataset does not support the Russian language.

One may also consider the idea of creating a topical dataset on the basis of a question-answering website. Creators of the `Yahoo!Answers` dataset [32] have implemented this idea. This 10-topic dataset contains questions and answers for topics from the "Yahoo Answers" service. However, the assortment of topics included in this dataset is far from exhaustive. This dataset also does not contain the Russian language.

The `MASSIVE` [9] dataset has been created for conversational topic and intent classification. In this dataset with 17k samples (train+test+valid), one of the 18 topic classes and one of the 60 intent classes is assigned to every utterance. This dataset is massively multilingual, as every utterance in this dataset is provided in 51 different languages (including Russian), adapted to the specifics of the corresponding countries. We note that this

dataset consists of conversational requests to a voice assistant. However, the nomenclature of topics provided in this dataset does not even remotely cover all possible user topics.

The nomenclature of topics covered in the dataset `DeepPavlov Topics` [26] is much larger, as 33 classes from this dataset cover a substantial number of possible conversational situations. However, this dataset does not cover the Russian language either.

The only publicly available Russian-language dataset we know that includes a significant number of conversational classes is `Chatbot-ru` [18]. This dataset has a very large nomenclature of Russian intents and topics (79 classes). However, the size of the dataset is far too small for such a large number of classes (∼7.1k total samples). In this dataset, intents are treated in the same way as topics, so the real number of topical classes and samples in this dataset is smaller. Given that this dataset is also imbalanced, a vast majority of topics in this dataset have less than 100 samples per class (or even much less, up to 10-20). Such a small number of samples per class makes the dataset suitable for the few-shot setting. However, it still leaves much room for improvement in terms of the dataset size expansion. Moreover, the variety of topical classes in this dataset is still incomplete and does not comprise some topics from [26].

As one can see from the above survey, not all topical datasets are suitable for use in a dialog system that works with real user phrases. Some datasets have too few classes, some other datasets have very domain-specific class nomenclature, and other datasets' examples are too different from real-world dialog data which can cause additional distortions. Furthermore, this field currently has a dire lack of topical datasets in Russian; existing Russian datasets are incomplete and either too small or too specific.

Knowledge transfer from the Russian language for topical datasets is also under-researched. Our work aims to bridge this gap.

§3. `RuQTopics` Dataset

In this work we examine `RuQTopics`, a Russian topic classification dataset. Raw data for this dataset was obtained from the "Yandex Que" question answering service raw data.[1]

[1] `https://huggingface.co/datasets/its5Q/yandex-q/blob/main/full.jsonl.gz`

Utterances in this dataset are labeled with 76 topics. We have selected the topics that we use below based on the dataset [26]. All utterances in the dataset contain questions. The questions in `RuQTopics` are short: 50% of the questions have less than 10 words and less than 1% have over 30 words. At the same time, answers in the dataset are mostly very long: only ≈1% of the answers have less than 10 words, and 50% of the answers have 65 words or less. 91.6% of the answers consist of 256 words or less.

The topic of every question corresponds to its section on "Yandex Que". For every question, we have selected the answer with the best quality score (or the first such answer if there are several). For some questions, the answer was empty.

We have split the question-answer pairs we obtained into two parts. In part 1 (single-label) we select only those pairs where the question belongs to only one topic, and the answer to this question either does not exist or can be found solely in this topic. All other examples belong to part 2 (multi-label). Here and below, we work only with the single-label part of the `RuQTopics`.

For all 76 topics, we have obtained 532,590 unique questions, of which 403,938 are answered. The single-label part of the dataset contains 361,650 questions, of which 266,597 are answered. The multi-label part of the dataset contains 170,930 questions, of which 137,431 are answered.

Additionally, we have selected the matched part of `RuQTopics` as a subset of the single-label one. If a question is answered, and the answer to this question can be found in only one topic (the same topic as the question has), the question-answer pair was included not only in the single-label part of the dataset but also in the matched part.

Table 1 shows the sizes of all parts of `RuQTopics` for every class we use in this work.

We note that, as some `RuQTopics` classes are similar to each other, to use this dataset in applications one might have to merge some of the classes.

For our experiments on this dataset, we have trained Transformer-based models with hyperparameters and backbones described in the next section.

§4. Experimental Setup

While training all models described in this work, we used the following hyperparameters: batch size 160, optimizer AdamW [12], betas (0.9,0.99), initial learning rate 2e-5, learning rate drops by 2 times if accuracy does

Table 1. `RuQTopics` sizes for different splits and all classes considered in this work.

data type	**single-label**		**multi-label**		**matched**
	all	answered	all	answered	
Full dataset size	361,650	266,597	170,930	137,341	264,786
6-class subset size	18864	15912	27191	20569	15830
music	9,514	5,809	4,456	3,287	5,797
food, drinks and cooking	5,750	4,758	14,096	11,084	4,723
media and communications	4,505	2,637	5,577	3,948	2,619
transport	2,435	1,625	1,933	1,387	1,613
news	945	602	912	720	600
weather	890	481	217	143	478

not improve for 2 epochs, validation patience 3 epochs, max 100 training epochs. The max sequence length is 256 tokens. We performed three random restarts for all experiments and averaged the metrics.

We performed the experiments on multiple backbones from the HuggingFace `Transformers` library [31], which all have a similar BERT-like architecture: *bert-base-multilingual-cased* [10], *DeepPavlov/distilrubert-tiny-cased-conversational* [13], *ai-forever/ruBert-base* [27] and *DeepPavlov/rubert-base-conversational-cased* [20]. The models *ai-forever/ruBert-base* and *DeepPavlov/rubert-base-conversational-cased* are similar, but they have a slightly different number of parameters because of different tokenization. We describe the difference between these backbones in Table 2.

4.1. Model Benchmarking. To benchmark the performance of models trained on our dataset on the conversational tasks, we utilized the `MASSIVE` dataset for evaluation. We have selected this dataset because it contains data manually checked by crowdsourced workers, and it consists of conversational utterances as well as `RuQTopics`.

While comparing our dataset with `MASSIVE`, we saw that only six `MASSIVE` classes can be directly mapped to `RuQTopics`. Therefore, we trained all described models only on the six corresponding classes from the single-label subset of `RuQTopics`: *food, drinks, and cooking* (corresponds to the *cooking* `MASSIVE` class), *news* (corresponds to the *news* `MASSIVE` class), *transport* (corresponds to the *transport* `MASSIVE` class), *music* (corresponds to the *music* `MASSIVE` class), *media and communication* (corresponds to the *social* `MASSIVE` class) and *weather* (corresponds to the

Table 2. Parameters of different backbone models considered in this work.

Backbone model	**Abbr.**	**Multilingual**	**Layers**	**Parameters**
DeepPavlov/distilrubert-tiny-cased-conversational [13]	*rubert-tiny*	no	2	107M
DeepPavlov/rubert-base-cased-conversational [20]	*rubert*	no	12	177.9M
bert-base-multilingual-cased [10]	*multbert*	yes	12	177.9M
ai-forever/ruBert-base [27]	*ru-sbert*	no	12	178.3M

weather `MASSIVE` class). We did not merge `RuQTopics` classes even though it could have additionally improved the results for *cooking* and *transport* `MASSIVE` classes.

We validated all models on the Russian `MASSIVE` validation 6-class subset and tested them on the concatenation of train and test 6-class subsets of `MASSIVE`. Here and below, we denote this subset concatenation as the "custom test set".

This method allows to test whether the dataset is suitable for conversational topic classification, at least on a subset of classes. However, since examples for all classes were collected similarly, we expect that other classes from the `RuQTopics` are as suitable for conversational topic classification as these six ones.

§5. Dataset Preprocessing

We needed to identify the best method of `RuQTopics` preprocessing for the best performance on conversational tasks. Specifically, we have compared five different methods of preprocessing for the `RuQTopics` dataset. We call them "modes" in Table 3. In these modes:

- **Q** means using only questions.
- **A** means using only answers.
- **Q [SEP] A** means using the concatenation of every question with the corresponding answer using the [SEP] token. If the question is unanswered, it means using only the question.

For all of these preprocessing methods, we performed training on the matched version of the `RuQTopics` (column "matched" in Table 1). This training mode allows to make a fair comparison between features obtained

by different preprocessing methods, as the number of training samples in this method is the same regardless of how we preprocess the data. We present in Table 3 the results obtained in this training mode. We also present in Table 4 the results obtained by training on the full single-label version of this dataset (column "single-label" from Table 1).

As one can see from Table 3, the question-only setting yields larger scores than the answer-only setting. This conclusion holds for all considered backbone models, proving that questions are the most informative feature in the `RuQTopics` dataset. If we concatenate questions with answers, the scores do not change significantly compared to the question-only setting.

We have also tried using answers that are summarized by TextRank [1] instead of the full answers in the experiments. The summarized answer-only setting has shown sustainably worse results than the answer-only one, and the concatenation of questions to summarized answers has given the same scores as the concatenation of questions to answers.

Overall, all Russian models show similar results, and the multilingual model expectedly trails behind them all.

All these conclusions are also valid for the full 6-class subset, as one can see from Table 4. For the experiments in the next sections, we chose the **Q** preprocessing mode, as all other modes are either more complicated and do not improve the results (**Q [SEP] A**) or show worse results (**A**).

§6. Evaluation for all `RuQTopics` classes

Another important task is to figure out how well the `RuQTopics` classes can be distinguished from each other. To do so, we perform 5-fold cross-validation on all questions from the single-label `RuQTopics` part. We present the results in Table 5.

The results could have been additionally improved by merging some classes from similar "Yandex.Que" topics. But even without that, Russian non-distilled backbones show an accuracy of 73.7-74.0%, whereas Russian distilled backbones fare slightly worse (72.2% accuracy). The multilingual backbone trails slightly behind these backbones by this measure (71.4% accuracy), as expected. This shows that topical classes in the dataset can be distinguished from each other with sufficiently high accuracy.

§7. Cross-Lingual Knowledge Transfer

After we had selected the best setting, the following questions emerged: how effectively does knowledge from this setting transfer across multiple

Table 3. Accuracy (F1) of the backbones (abbreviated as in Table 2) on the Russian `MASSIVE` custom test set, trained on `RuQTopics` 6-class **matched** subsets with different preprocessing modes (Section 5); avg over 3 runs.

Model	**Mode**		**Total**		**music**		**cooking**		**news**		**transport**		**weather**		**social**	
			Acc	Mc-F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
ru	**Q**		84.2	83.4	94.3	87.5	99.0	82.0	80.1	83.1	92.3	90.7	81.3	89.0	65.5	68.1
rutiny	**Q**		85.7	84.9	94.3	89.1	99.2	81.5	76.8	83.3	93.8	92.4	84.2	90.4	73.2	72.7
rusber	**Q**		85.2	84.1	93.9	90.2	98.9	79.8	80.3	84.3	92.8	90.6	85.7	90.9	64.9	69.0
mult	**Q**		79.1	77.7	93.2	79.8	98.0	74.0	73.5	80.3	89.9	87.9	74.4	83.9	55.2	60.1
ru	**A**		80.7	79.1	95.8	78.5	98.2	81.6	66.3	77.2	91.5	87.9	87.1	91.1	51.6	58.3
rutiny	**A**		82.4	81.0	96.4	81.1	99.3	78.1	71.3	80.8	88.9	90.3	86.7	90.8	60.0	64.8
rusber	**A**		82.3	80.7	94.9	81.2	98.9	80.6	72.6	80.8	89.8	89.2	88.7	91.2	54.5	61.3
mult	**A**		76.8	75.3	94.3	76.6	96.1	70.0	68.3	77.5	83.0	83.4	78.6	85.2	50.8	59.0
ru	**Q**	**[SEP] A**	85.7	85.2	92.3	90.1	97.4	86.4	79.1	82.9	93.3	91.5	86.4	91.1	70.3	69.5
rutiny	**Q**	**[SEP] A**	85.0	84.2	95.3	87.2	98.3	82.4	75.5	82.3	89.7	92.0	86.4	91.1	72.4	70.5
rusber	**Q**	**[SEP] A**	85.3	84.7	92.7	89.5	98.7	85.8	80.7	82.1	91.0	91.1	87.1	92.0	66.7	67.9
mult	**Q**	**[SEP] A**	78.5	77.6	93.0	82.5	95.9	72.9	67.0	77.1	86.1	85.6	75.1	84.0	65.1	63.5

Table 4. Accuracy (F1) of the backbones (abbreviated as in Table 2) on the Russian `MASSIVE` custom test set, trained on `RuQTopics` 6-class **full** subsets with different preprocessing modes (Section 5); averaged over 3 runs.

Model	**Mode**	**Total**		**music**		**cooking**		**news**		**transport**		**weather**		**social**	
		Acc	Mc-F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
ru	**Q**	85.0	84.3	94.7	87.5	98.4	86.0	82.5	82.7	92.1	92.1	82.5	89.6	66.1	68.0
rutiny	**Q**	85.7	85.2	95.0	87.5	98.7	87.3	82.2	82.9	92.8	92.3	84.3	90.6	67.3	70.3
rusber	**Q**	85.5	84.9	93.7	89.9	98.8	87.2	83.0	83.1	93.1	91.1	85.1	91.1	64.3	67.2
mult	**Q**	80.8	79.8	94.3	77.3	97.2	82.8	75.5	81.1	90.0	90.5	78.5	86.0	57.5	61.0
ru	**Q [SEP] A**	85.4	84.9	94.0	88.5	97.6	87.1	82.5	83.7	93.1	91.5	83.6	90.0	67.1	68.8
rutiny	**Q [SEP] A**	85.3	84.7	94.3	87.4	97.9	86.4	79.3	81.4	91.6	92.8	86.0	91.0	68.2	69.1
rusber	**Q [SEP] A**	85.1	84.2	93.1	91.6	98.3	88.4	88.1	81.2	93.3	91.9	86.2	91.6	53.7	60.7
mult	**Q [SEP] A**	80.0	79.7	94.2	77.0	95.1	85.5	74.9	80.3	87.8	88.0	73.8	83.8	64.6	63.6

Table 5. Accuracy (Macro-F1) of different backbone models for the 5-fold cross-validation on all questions from the single-label part of the `RuQTopics` dataset (76 classes). Backbones are abbreviated as in Table 2.

Model	Average		Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
	Acc	Mc-F1	Acc	Mc-F1	Acc	Mc-F1	Acc	Mc-F1	Acc	Mc-F1	Acc	Mc-F1
rusber	74.0	53.4	73.7	54.3	73.8	52.8	73.9	53.0	74.1	54.2	74.2	52.9
ru	73.7	52.5	73.5	52.9	73.7	51.9	73.6	52.3	73.9	53.1	73.9	52.3
rutiny	72.2	50.9	72.0	49.7	72.2	50.9	72.0	51.4	72.4	51.1	72.3	51.6
mult	71.4	51.9	71.2	52.4	71.5	51.9	71.5	51.4	71.2	51.6	71.7	52.1

languages? What influences the efficiency of this transfer? To answer these questions, we pre-trained *bert-base-multilingual-cased*, which allows effective cross-lingual transfer learning on different NLP tasks [7, 15], on the data from **full** validation 6-class `RuQTopics` subset, which are preprocessed by the **Q** preprocessing mode. For this backbone, using the full subset instead of the matched subset gave 1-2% growth in accuracy and macro-F1 for the Russian language.

In this stage, we apply this model not only on the Russian `MASSIVE` but also on all other languages it contains.[2]

An interesting research question is the correlation of model quality for different languages with the pretraining sample size for that language. The authors of *bert-base-multilingual-cased* claim [10] that the learning sample for every utilized language was comprised of Wikipedia texts for that language and that they performed an exponential smoothing of the training sample with the factor of 0.7 to balance the languages. Therefore, as a proxy of the Wikipedia size for every language, we used the number of articles in the Wikipedia of this language at the time of the BERT article's release, smoothed by the factor of 0.7.

We present the metrics obtained by the evaluation of the multilingual BERT on the custom `MASSIVE` test subset for all languages in Tables 6 and 7. For every language, we also provide its genealogical distance to Russian (calculated as in [3]) and the original Wikipedia size we used in the same table.

The Spearman correlation of the total accuracy with the smoothed Wikipedia size is 0.773 (p-value 2.997e-11, 95% CI: [0.63, 0.86]). At the

[2]We use `MASSIVE` version 1.1, which contains the Catalan language. For the Chinese language, we have utilized both sets of characters as `MASSIVE` has two Chinese versions.

Table 6. Accuracy (F1) of the *bert-base-multilingual-cased* on the custom test set for all `MASSIVE` languages. The model was trained on the **Q** version of **full** `RuQTopics` 6-class subset and validated on the 6-class validation set of Russian `MASSIVE`. **Code** means ISO 639-1 language code, **Dist** means genealogical distance between that language and Russian [3]. **N** means the number of Wikipedia articles in that language as of 11-10-2018. We trained on the **full** single-label version of `RuQTopics`. Averaged over three runs.

Language	**Code**	**Dist**	**N**	**Metrics**	
				Acc	Mc-F1
Russian	ru	0	1,501,878	80.8	79.8
Chinese-TW	zh-TW	92.2	1,025,366	79.6	79.1
Chinese	zh	92.2	1,025,366	78.0	77.7
English	en	60.3	5,731,625	75.2	75.6
Japanese	ja	93.3	1,124,097	72.4	70.5
Slovenian	sl	4.2	162,453	70.3	69.0
Swedish	sv	59.5	3,763,579	70.2	69.6
Malay	ms	n/c	320,631	68.9	67.7
Italian	it	45.8	1,466,064	68.8	68.0
Indonesian	id	91.2	440,952	68.7	67.5
Dutch	nl	64.6	1,944,129	68.7	68.5
Portuguese	pt	61.6	1,007,323	68.6	68.7
Spanish	es	51.7	1,480,965	68.2	68.0
Danish	da	66.2	240,436	67.8	66.7
French	fr	61.0	2,046,793	65.5	65.5
Persian	fa	72.4	643,750	65.2	64.2
Turkish	tr	86.2	316,969	64.5	62.4
Vietnamese	vi	95.0	1,190,187	64.3	65.1
Norwegian B	nb	67.2	495,395	64.3	64.0
Polish	pl	5.1	1,303,297	64.2	62.2
Azerbaijani	az	87.7	138,538	63.9	63.1
Catalan	ca	60.3	591,783	61.4	60.4
Hungarian	hu	87.2	437,984	61.3	60.0
Hebrew	he	88.9	231,868	60.9	59.5
Hindi	hi	69.8	127,044	60.7	58.7

Table 7. Table 6, continued.

Language	Code	Dist	N	Metrics	
				Acc	Mc-F1
Korean	ko	89.5	429,369	60.4	59.6
Romanian	ro	55.0	388,896	57.1	53.9
Urdu	ur	66.7	140,939	56.4	55.9
Arabic	ar	86.5	619,692	56.2	55.7
Kannada	kn	90.8	23,844	56.1	53.0
Filipino	tl	91.9	80,992	55.0	51.3
Telugu	te	96.7	69,354	53.7	49.3
Finnish	fi	88.9	445,606	53.3	51.3
Burmese	my	86.0	39,823	52.5	49.7
Afrikaans	af	64.8	62,963	52.4	50.3
Tamil	ta	94.7	118,119	52.4	50.1
German	de	64.5	2,227,483	52.2	51.6
Albanian	sq	69.4	74,871	51.5	47.2
Latvian	lv	49.1	88,189	49.6	48.4
Malayalam	ml	96.7	59,305	48.7	46.3
Armenian	hy	77.8	246,571	48.1	47.5
Bangla	bn	66.3	61,294	47.3	45.3
Thai	th	89.5	127,010	46.5	44.9
Greek	el	75.3	153,855	46.3	44.8
Georgian	ka	96.0	124,694	39.2	38.1
Javanese	jv	95.4	54,964	38.7	37.1
Mongolian	mn	86.2	18,353	36.6	33.7
Icelandic	is	68.9	45,873	32.6	29.9
Swahili	sw	95.1	45,806	31.0	28.0
Welsh	cy	75.5	101,472	28.5	25.3
Khmer	km	97.1	6,741	16.1	8.6
Amharic	am	86.6	14,375	12.1	5.0

same time, the Spearman correlation of the total accuracy with the genealogical distance to the Russian is -0.323 (p-value 0.022, 95% CI: [-0.55, -0.05]).[3] If we take into account the smoothed Wikipedia size as the confounding variable, the partial correlation of the total accuracy with the

[3]We excluded the Russian language itself from the calculations.

genealogical distance to the Russian becomes -0.027 (p-value 0.856, 95% CI: [-0.31, 0.26]), which is statistically insignificant.

§8. Discussion

As one can see, the `RuQTopics` dataset overall is well suited for conversational topic classification.

We hypothesize that, apart from topical classification, this dataset can also be utilized for the question-answering task. However, we leave testing this hypothesis for future research.

In the case of question classification, different Russian-only baseline models trained on the `RuQTopics` 6-class subset obtain an accuracy of around 85% on the subset of the same six classes from the Russian `MASSIVE` (Table 4).

We obtain such accuracy only if we utilize questions from `RuQTopics` in the training features (either by themselves or in concatenation with answers), which proves that the questions are the most informative features in this dataset.

Surprisingly, switching between different Russian-only baseline models, including even the two-layer distilled one, did not significantly alter the results. That proves that the distilled conversational models are well suited for conversational tasks, especially in the case of constrained computational resources.

For training models on all 76 classes of the `RuQTopics` in the question-only setting, all backbones show accuracy above 70%. That shows that the dataset is suitable for the classification task as a whole, not just as a six-class subset.

In the case of evaluation of the multilingual BERT (trained on the `RuQTopics` question subset) on all languages included in the `MASSIVE` dataset, the accuracy by language closely correlates with the approximated size of the BERT pretraining dataset for that language (Spearman correlation 0.773 with p-value 2.997e-11). We have approximated the dataset size by exponentiation of the language-wise number of Wikipedia size as of 11-10-2018 (date of release of the [8]) by 0.7, similarly to the original article.

Such correlation was obtained even though an average Wikipedia article in different languages has a different number of tokens and sentences. We suppose that if we had the precise number of training samples for every language that the original multilingual model received at the pretraining

stage, the correlation would have been even higher; however, the authors of the original BERT article provided neither the original training sample nor its language-wise size.

At the same time, the correlation of model scores with the genealogical distance to the Russian is statistically insignificant. This leads to the conclusion that the main factor determining the quality of knowledge transfer between different languages in the multilingual BERT-like models is, by far, the size of the pretraining sample for this language. We can suppose that for the case of languages that are very linguistically close (e.g. Russian and Belarusian) such closeness also impacts knowledge transfer, but examining the importance of this factor requires additional research.

§9. Conclusion

In this work, we have investigated knowledge transfer from the `RuQTopics` dataset. This Russian topical dataset combines a large sample number (361,560 single-label, 170,930 multi-label) with extensive class coverage (76 classes). We have prepared this dataset from the "Yandex Que" raw data.

By evaluating the `RuQTopics`-trained models on the six matching classes of the Russian `MASSIVE` subset, we have proved that the `RuQTopics` dataset is suitable for real-world conversational tasks, as Russian-only models trained on this dataset consistently yield the accuracy around 85% on this subset (Table 4). We also have figured out that for the multilingual BERT, trained on the `RuQTopics` and evaluated on the same six classes of `MASSIVE` (for all `MASSIVE` languages), language-wise accuracy closely correlates with the approximate size of the pretraining BERT's data for the corresponding language. At the same time, the correlation of the language-wise accuracy with the genealogical distance from the Russian is not statistically significant.

§10. Acknowledgments

We are grateful to Pavel Levchuk for the raw data collection, to Anastasiya Chizhikova for her help with the English language, and to Vasily Konovalov and Alexander Popov for valuable remarks.

References

1. F. Barrios, F. López, L. Argerich, R. Wachenchauzer, *Variations of the similarity function of textrank for automated summarization*, CoRR **abs/1602.03606** (2016).

2. D. Baymurzina, D. Kuznetsov, D. Evseev, D. Karpov, A. Sagirova, A. Peganov, F. Ignatov, E. Ermakova, D. Cherniavskii, S. Kumeyko, O. Serikov, Y. Kuratov, L. Ostyakova, D. Kornev, M. Burtsev, *Dream technical report for the alexa prize 4*, Alexa Prize SocialBot Grand Challenge 4 Proceedings (2021).
3. V. Beaufils, J. Tomin, *Stochastic approach to worldwide language classification: the signals and the noise towards long-range exploration*, (2020), Implementation we used: `http://www.elinguistics.net/Compare\_Languages.aspx`.
4. P. Blinov, *Dataset of russian reviews about medical facilities*, `https://huggingface.co/datasets/blinoff/healthcare\_facilities\_reviews`, 2022, Accessed: 2023-02-17.
5. V. Morris, D. van Strien, G. Tolfo, L. Afric, S. Robertson, P. Tiney, A. Dogterom, I. Wollner, *19th century books - metadata with additional crowdsourced annotations*, 2021.
6. I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. M. Katz, N. Aletras, *Lexglue: A benchmark dataset for legal language understanding in english*, Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Dubln, Ireland), 2022.
7. A. Chizhikova, V. Konovalov, M. Burtsev, *Multilingual case-insensitive named entity recognition*, Advances in Neural Computation, Machine Learning, and Cognitive Research VI (Cham) (Boris Kryzhanovsky, Witali Dunin-Barkowski, Vladimir Redko, and Yury Tiumentsev, eds.), Springer International Publishing, 2023, pp. 448–454.
8. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, *BERT: Pre-training of deep bidirectional transformers for language understanding*, 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), 2019, p. 4171:4186 (english).
9. J. FitzGerald, C. Hench, C. Peris, S. Mackie, K. Rottmann, A. Sanchez, A. Nash, L. Urbach, V. Kakarala, R. Singh, S. Ranganath, L. Crist, M. Britan, W. Leeuwis, G. Tur, P. Natarajan, *Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages*, 2022.
10. S. Petrov, J. Devlin, *Official description of the multilingual bert models from google research*, `https://github.com/google-research/bert/blob/master/multilingual.md`, 2019.
11. P. Keung, Y. Lu, G. Szarvas, N. A. Smith, *The multilingual amazon reviews corpus*, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020.
12. D. P. Kingma, J. Ba, *Adam: A method for stochastic optimization*, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (Yoshua Bengio and Yann LeCun, eds.), 2015.
13. A. Kolesnikova, Y. Kuratov, V. Konovalov, M. Burtsev, *Knowledge distillation of russian language models with reduction of vocabulary*, 2022.
14. V. Konovalov, R. Artstein, O. Melamud, I. Dagan, *The negochat corpus of human-agent negotiation dialogues*, Proceedings of the Tenth International Conference on

Language Resources and Evaluation (LREC'16) (Portorož, Slovenia), European Language Resources Association (ELRA), May 2016, pp. 3141–3145.
15. V. Konovalov, P. Gulyaev, A. Sorokin, Y. Kuratov, M. Burtsev, *Exploring the bert cross-lingual transfer for reading comprehension*, Komp'juternaja Lingvistika i Intellektual'nye Tehnologiithis, 2020, pp. 445–453.
16. V. Konovalov, O. Melamud, R. Artstein, I. Dagan, *Collecting Better Training Data using Biased Agent Policies in Negotiation Dialogues*, Proceedings of WOCHAT, the Second Workshop on Chatbots and Conversational Agent Technologies (Los Angeles), Zerotype, September 2016.
17. V. P. Konovalov, Z. B. Tumunbayarova, *Learning word embeddings for low resource languages: the case of buryat*, Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, 2018, pp. 331–341.
18. I. Koziev, *Chatbot-ru: Russian intent and topic classification dataset*, `https://github.com/Koziev/chatbot/blob/master/data/intents.txt`, 2020.
19. Y. M. Kuratov, I. F. Yusupov, D. R. Baymurzina, D. P. Kuznetsov, D. V. Cherniavskii, A. Dmitrievskiy, E. S. Ermakova, F. S. Ignatov, D. A. Karpov, D. A. Kornev, T. A. Le, P. Y. Pugin, M. S. Burtsev, *Socialbot dream in alexa prize challenge 2019*, Proceedings of Moscow Institute of Physics and Technology **13** (2021), no. 3, 62–89.
20. Y. Kuratov, M. Y. Arkhipov, *Adaptation of deep bidirectional multilingual transformers for russian language*, CoRR **abs/1905.07213** (2019).
21. Y. Kuratov, I. Yusupov, D. Baymurzina, D. Kuznetsov, D. Cherniavskii, A. Dmitrievskiy, E. Ermakova, F. Ignatov, D. Karpov, D. Kornev, and Others, *Dream technical report for the alexa prize 2019*, 3rd Proceedings of Alexa Prize (2019).
22. J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer, C. Bizer, *Dbpedia – a large-scale, multilingual knowledge base extracted from wikipedia*, Semantic Web Journal **6** (2014).
23. Y. Liang, N. Duan, Y. Gong, N. Wu, F. Guo, W. Qi, M. Gong, L. Shou, D. Jiang, G. Cao, X. Fan, R. Zhang, R. Agrawal, E. Cui, S. Wei, T. Bharti, Y. Qiao, J.-H. Chen, W. Wu, S. Liu, F. Yang, D. Campos, R. Majumder, and M. Zhou, *XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation*, arXiv **abs/2004.01401** (2020).
24. J. Niklaus, V. Matoshi, P. Rani, A. Galassi, M. Stürmer, I. Chalkidis, *Lextreme: A multi-lingual and multi-task benchmark for the legal domain*, 2023.
25. A. Perevalov, *Pstu dataset: classification of university-related topics*, `https://github.com/Perevalov/pstu\_assistant/blob/master/data/data.txt`, 2018.
26. B. Sagyndyk, D. Baymurzina, M. Burtsev, *DeepPavlov topics: Topic classification dataset for conversational domain in English*, Advances in Neural Computation, Machine Learning, and Cognitive Research VI (Cham) (Boris Kryzhanovsky, Witali Dunin-Barkowski, Vladimir Redko, and Yury Tiumentsev, eds.), Springer International Publishing, 2023, pp. 371–380.
27. SberDevices, *rut5, ruroberta, rubert: How we trained a series of models for the russian-language*, `https://habr.com/ru/company/sberbank/blog/567776/`, 2021, HuggingFace model link: `https://huggingface.co/sberbank-ai/ruBert-base`. Accessed: 2023-02-17.
28. T. Scialom, P.-A. Dray, S. Lamprier, B. Piwowarski, and J. Staiano, *Mlsum: The multilingual summarization corpus*, arXiv preprint arXiv:2004.14900 (2020).

29. E. Stamatatos, *On the robustness of authorship attribution based on character n-gram features*, Journal of Law and Policy **21** (2013), 421–439.
30. M. Suzgun, L. Melas-Kyriazi, S. K. Sarkar, S. D. Kominers, S. M. Shieber, *The harvard uspto patent dataset: A large-scale, well-structured, and multi-purpose corpus of patent applications*, arXiv preprint arXiv:2207.04043 (2022).
31. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, Teven Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, *Transformers: State-of-the-art natural language processing*, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (Online), Association for Computational Linguistics, October 2020, pp. 38–45.
32. X. Zhang, J. J. Zhao, Y. LeCun, *Character-level convolutional networks for text classification*, NIPS, 2015.
33. L. Zhou, J. Gao, D. Li, H.-Y. Shum, *The design and implementation of xiaoice, an empathetic social chatbot*, 2018.

Moscow Institute of Physics
and Technology,
Dolgoprudny, Russia
E-mail: dmitrii.a.karpov@phystech.edu

London Institute for Mathematical Sciences,
London, United Kingdom
E-mail: mbur@lims.ac.uk

Поступило 6 сентября 2023 г.