National Research University Higher School of Economics

as a manuscript

Maria Tikhonova

Language Model Evaluation in Natural Language Understanding

PhD Dissertation Summary

for the purpose of obtaining
academic degree Doctor of Philosophy
in Computer Science

Moscow – 2023

The PhD dissertation was prepared at National Research University Higher School of Economics

Academic Supervisor: Konstantin Vorontsov, Doctor of Physico-Mathematical Sciences, Professor of Russian Academy of Sciences, Deputy Head of the Department of Mathematical Methods of Forecasting, CMC MSU, Head of the Department of Machine Learning and Digital Humanities, MIPT, Professor, Machine Intelligence Laboratory, MIPT, senior researcher, Department "Intelligent Systems", Computing Centre of Russian Academy of Sciences

Academic Supervisor: Tatiana Shavrina, PhD in Computational Linguistics, Director of Key research projects in NLP, AIRI, Senior Researcher, Institute of Linguistics, RAS

1. Dissertation Topic

Problem Statement and Motivation

The task of language modeling has received considerable attention in recent years. Nowadays, language models are the core part of *Natural Language Processing (or NLP)*. Language model applications are numerous; they are used practically in all NLP tasks. Among such tasks are text classification (sentiment analysis, spam detection, genre classification, response classification, etc.), information retrieval, information extraction (named entity recognition, part of speech tagging, etc.), a wide range of text generation tasks (text summarization, question answering task, machine translation, rewriting, etc.) and many others.

The predominant approach in language modeling nowadays is neural-based. Today, there exist a large variety of language models that differ in architecture, the data they were trained on, parameter setup used during training, and other tiny details. This immense diversity of language models raises the question of their efficiency and how well they understand natural language.

Thus, the following questions connected with language model evaluation become increasingly relevant:

1) there is a need to develop methods for the quantitative evaluation of language models in various NLP tasks;
2) there is a need to develop reliable test systems, procedures, and tools that can be used to evaluate certain aspects of language models and compare them with each other.

This thesis focuses on one of the aspects of language model evaluation. Namely, it is devoted to language model evaluation methods in *Natural Language Understanding (or NLU)*.

Aims of the Thesis Research

The main goal of this Thesis is to develop methods for language model evaluation in Natural Language Understanding and to create the necessary set of tests and tools to perform this evaluation. To achieve this goal, the following tasks are set.

1. Develop a method for systematic language model evaluation in Natural Language Understanding tasks.
2. Develop a method for language model stability evaluation, on the basis of Natural Language Inference task.
3. Conduct a series of experiments to evaluate the stability of the Multilingual BERT [Delvin J. et al., 2019] model in several languages. Test the hypothesis about the influence of the training dataset size on the stability of the model results. Conduct a comparative analysis between different languages. These experiments should be conducted using the method from item 2.

Related Work

This section describes scientific research conducted by the time the thesis research began. It is divided into parts corresponding to the tasks from the section "Aims of the Thesis Research".

Develop a system for language model evaluation in Natural Language Understanding tasks

Today, the benchmark approach is commonly recognized as the standard way of language model evaluation in Natural Language Understanding. A benchmark consists of several tasks (or tests), where each task tests a particular NLU aspect. To comprehensively evaluate the language model, it has to solve all the tasks. In recent years several NLU benchmarks have been introduced. SentEval [Conneau et al., 2018a] is one of the first frameworks that evaluates sentence embedding quality.

The *General Language Understanding Evaluation (or GLUE)* benchmark [Wang et al., 2018] is a collection of tools for evaluating language model performance across a diverse set of NLU tasks. Today GLUE is commonly recognized as a standard benchmark. The idea of GLUE is further developed in the SuperGLUE [Wang, Alex, et al., 2019] benchmark, which follows the GLUE paradigm for evaluating language models based on NLU tasks. Compared to GLUE, SuperGLUE includes more complex tasks, some of which require reasoning capabilities or are aimed at detecting ethical biases. A few recent studies [Kovaleva et al., 2019; Warstadt et al., 2019] suggest that GLUE tasks may not be sophisticated enough and do not require much task-specific linguistic knowledge. Therefore, SuperGLUE, which is more challenging, becomes preferable for language model evaluation.

There exist several GLUE and SuperGLUE analogs in other languages: FGLUE [Le H. et al., 2019], KLEJ [Rybak P. et al., 2020], and CLUE [Xu L. et al., 2020] – French, Polish, and Chinese versions of the benchmark respectively. And several multilingual benchmarks, such as XGLUE [Liang Y. et al., 2020] and XTREME [Hu J. et al., 2020], aimed at language model evaluation in several languages simultaneously.

However, most of the current research in this area focuses on the English language. It presents benchmarks and collections of tasks specifically for this language, while Russian, which appears only in a few multilingual benchmarks, is underrepresented. At the beginning of the thesis research, there was no system of tests for the comprehensive assessment of the language model abilities in Natural Language Understanding similar to GLUE or SuperGLUE for the English language.

Develop a method for language model stability evaluation in Natural Language Inference task and conduct a series of experiments to evaluate the stability of Multilingual BERT

Natural Language Inference (or NLI) [Storks S. et al., 2019] task has received considerable attention in recent years, and several NLI datasets have been proposed. Among them are RTE [Dagan I. et al., 2005], SICK [Marelli M. et al., 2014], SNLI [Bowman S. R. et al., 2015], MNLI [Williams A. et al., 2017], and XNLI [Conneau et al., 2018b]. In addition, the diagnostic dataset from the GLUE [Wang et al., 2018] benchmark is worth mentioning. Today it is recognized as a standard dataset for examining linguistic knowledge of language models in English for the NLI task.

Several recent studies have examined the role of optimization, training data, and implementation choices on the stability of language models [Henderson P. et al., 2018; Madhyastha P. et Jain R., 2019; Dodge J. et al., 2020]. In the experiments [Devlin J. et al., 2019] BERT has demonstrated unstable behavior when fine-tuned on small datasets across multiple restarts. Research [Lee C. et al., 2019; Mosbach M. et al., 2020; Hua H. et al., 2021] shows that changing random seeds in the process of fine-tuning can lead to significant variations of the validation performance in various NLP tasks, including the tasks from the GLUE benchmark.

Several works covered in the survey [Rogers A. et al., 2020] are devoted to the linguistic analysis of the BERT model and the influence of fine-tuning on the model knowledge. Current research studies various linguistic phenomena, including syntactic properties [Warstadt A. et Bowman S., 2019], semantic knowledge [Goldberg Y., 2019], common sense [Cui L. et al., 2020], and others [Ettinger A., 2020].

In light of the fact that the research mentioned above highlights the unstable, predominantly random behaviour of the BERT model, the development of methods for language model stability evaluation becomes highly relevant, as well as the study of the linguistic abilities of BERT.

This thesis continues research in this area, considering the stability of the BERT model in the context of learning certain linguistic features for the NLI task. Namely, it analyzes the stability of the *multilingual BERT model (or mBERT)* with respect to diagnostic inference features and, therefore, extends the experimental setup to the multilingual setting.

Novelty of the Research

1. *A novel method* for the language model stability evaluation in Natural Language Inference task is proposed.

2. Based on the method from point 1, *a methodology for the multilingual language model evaluation in five languages* is proposed.

3. *A unique study of the stability of the multilingual BERT model for the NLI task in five languages* is conducted, which reveals the relationship between the volume of training data and the model stability.

4. As a part of the development of the first Russian Natural Language Understanding benchmark, *a framework for language model evaluation* on this benchmark is developed. This framework is used in the original study to evaluate several pre-trained BERT models for the Russian language.

2. Key Results

Key Ideas to be Defended:

1. As a part of the development of the *Russian SuperGLUE[1] (RSG) benchmark*, a set of tests for comprehensive language model evaluation in Natural Language Understanding, a framework *jiant-russian* for language model evaluation on this benchmark is developed. This software can be used to evaluate language models from the *HuggingFace* project on Russian SuperGLUE tasks. This framework makes it possible to fix experimental design and setup for model evaluation and, as a result, ensure the reproducibility of the experiments. Therefore, *jiant-russian*, combined with the Russian SuperGLUE benchmark, presents a convenient tool for language model evaluation and their comparison that determines the practical significance of the framework. With the use of this framework, a series of experiments on language model evolution for the Russian language is conducted. Results are published in [Shavrina T. et al., 2020; Fenogenova A. et al., 2021], and the framework *jiant-russian* is available in the Russian SuperGLUE repository[2].
2. A method for the language model stability evaluation in Natural Language Inference task is proposed. Namely, the method makes it possible to evaluate how stable a language model learns linguistic features in the process of solving the NLI task. This result has both theoretical and methodological significance for language model evaluation. A detailed description of the method and the obtained results is published in [Tikhonova M. et al., 2022].
3. A study on the stability of the linguistic generalization abilities of the mBERT model in the NLI task is conducted. Namely, a detailed analysis of the impact of the random initialization and the volume of training data is performed. In the experiments, it is shown that the training

[1] https://russiansuperglue.com/
[2] https://github.com/RussianNLP/RussianSuperGLUE

dataset size in the standard benchmarks is insufficient for the model stability with respect to learning various linguistic features. However, this stability can be improved for all languages by using additional data. The usage of extra data leads to a score improvement of 49% and model stability increase by 64%. A detailed description of the experiments is published in [Tikhonova M. et al. 2022].

Personal Contribution to the Ideas to be Defended

In [Tikhonova M. et al., 2022], the author proposes a language model stability evaluation method. Using this method, the author conducts a series of experiments on the stability of the multilingual BERT model in the NLI task in five languages and examines the influence of the additional training data on model stability and its overall score.

As a part of developing the *Russian SuperGLUE benchmark*, aimed at the comprehensive language model evaluation in Natural Language Understanding, in [Shavrina T. et al., 2020], the author develops a framework *jiant-russian* for the evaluation of language models from the HuggingFace project on this benchmark. With the use of *jiant-russian* the author conducts a series of experiments evaluating several language models on Russian SuperGLUE. As a continuation of this research, in [Fenogenova A. et al., 2021], the author improves the framework, adapts it to the changes in the set of tests, which were made as a part of this work, and includes the support of the new transformer-based models from HuggingFace project. In addition, the author compares several different pre-trained BERT models on Russian SuperGLUE, evaluating their abilities in Natural Language Understanding.

3. Publications and Approbation of Research

First-tier publications

1. [Tikhonova M. et al., 2022] **Tikhonova M.**, Mikhailov, V., Pisarevskaya, D., Malykh, V., Shavrina, T. Ad Astra or astray: Exploring linguistic knowledge of multilingual BERT through NLI task //Natural Language Engineering. – 2022. – C. 1-30. **Scopus, Q1**
2. [Shavrina T. et al., 2020] Shavrina T., Fenogenova A., Emelyanov A., Shevelev D., Artemova E., Malykh V., Mikhailov V., **Tikhonova M.**, Chertok A., Evlampiev A. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020. P. **(Core A)**

Second-tier publications

1. [Fenogenova A. et al., 2021] Alena Fenogenova, Tatiana Shavrina, Alexandr Kukushkin, **Maria Tikhonova**, Anton Emelyanov, Valentin Malykh, Vladislav Mikhailov, Denis Shevelev, Ekaterina Artemova. Russian SuperGLUE 1.1: Revising the Lessons not Learned by Russian NLP-models (2021) A Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2021” Moscow, 2021, Scopus
2. [Tikhonova. et al., 2021] **Tikhonova M**., Pisarevskaya D., Shavrina T., Shliazhko O. Using Generative Pretrained Transformer-3 Models for Russian News Clustering and Title Generation tasks. A Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2021” Moscow, 2021, Scopus
3. [Konodyuk N. et Tikhonova M., 2022] Konodyuk N., **Tikhonova M**. Continuous Prompt Tuning for Russian: How to Learn Prompts Efficiently with RuGPT3? //International Conference on Analysis of Images, Social Networks and Texts. – Springer, Cham, 2022. – C. 30-40., Scopus

Reports at conferences and seminars

1. The 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020, November 2020. Presentation: RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark, link **(core A conference)**
2. DIALOGUE Conference 2021 Presentation: Russian SuperGLUE 1.1: Revising the Lessons not Learned by Russian NLP-models, link
3. DIALOGUE Conference2021, Presentation: Using Generative Pretrained Transformer-3 Models for Russian News Clustering and Title Generation tasks, link
4. AIST Conference 2021 Presentation: Continuous Prompt Tuning for Russian: How to Learn Prompts Efficiently with RuGPT3, link
5. Artificial Intelligence and Natural Language Conference (AINL) 2022 Presentation: Multilingual GPT-3: downstream task evaluation with seq2seq setup, few-shot and zero-shot, link
6. Artificial Intelligence and Natural Language Conference (AINL) 2022 Presentation: Continuous prompt tuning for Russian: efficient solution for a variety of NLP task, link

4. CONTENTS

4.1 Language Modeling

The object of the research of this thesis is language models. Formally, a language model is a probability distribution over word sequences (or other text units, which are further referred to as tokens, units such as n-grams, characters, etc.). Thus, for every sequence of tokens$(w_1, \dots, w_n)$, a language model estimates the probability $P(w_1, \dots, w_n)$ of encountering such a sequence of tokens in the natural language. This research focuses on neural network language models based on the Transformer architecture [Vaswani A. et al., 2017]. This architecture is widely used in language modeling. Many transformer-based models have been proposed in recent years, including BERT [Delvin J. et al., 2019] and GPT-3 [Brown T. et al., 2020], which are studied in this research.

4.2 Russian SuperGLUE Benchmark

4.2.1 Russian SuperGLUE Tasks Description

As part of the thesis research, in [Shavrina T. et al., 2020, Fenogenova A. et al., 2021], a set of nine tasks is proposed. This set of tasks forms a benchmark called **Russian SuperGLUE** (RSG) for *Natural Language Understanding (NLU)* evaluation in Russian. RSG tasks test various NLU aspects and can be divided into six categories: Natural Language Inference (*TERRa, RCB),* Common Sense (*PARus, RUSSE*)*,* World Knowledge (*DaNetQA*), Machine Reading (*MuSeRC, RuCos*), Logical Reasoning (*RWSD*) and a diagnostic dataset *LiDiRus* additionally annotated with 33 linguistic phenomena under four high-level categories. Below is a brief description of each task, and aggregated information about datasets, their sizes, and scoring metrics is presented in Table 1.

Table 1. General information about Russian SuperGLUE tasks. Train/Val/Test stands for the example number int the train/validation/test sets, respectively. MCC = Matthew's correlation coefficient, EM = exact match.

Task	Task Type	Task Metric	Train	Val	Test
TERRa	NLI	Accuracy	2616	307	3198
RCB	NLI	Avg. Fl / Accuracy	438	220	438
LiDiRus	NLI & diagnostics	MCC	0	0	1104
RUSSE	Common Sense	Accuracy	19845	8508	18892
PARus	Common Sense	Accuracy	400	100	500
DaNet QA	World Knowledge	Accuracy	1749	821	805
MuSeRC	Machine Reading	Fl *I* EM	500	100	322
RuCoS	Machine Reading	Fl *I* EM	72193	7 577	7257
RWSD	Logical Reasoning	Accuracy	606	204	154

TERRA is a task aimed to capture textual entailment in a binary classification form. Given two text fragments (*premise* and *hypothesis*), the task is to recognize whether the meaning of one text can be inferred from the other.

Terra task example:

Premise: *Автор поста написал в комментарии, что провалилась канализация.*

Hypothesis: *Автор поста написал про канализацию.*

Label: *Entailment*

RCB is a Natural Language Inference task in the form of three class classification (*entailment, contradiction, neutral*).

RCB task example:

Text: *Сумма ущерба составила одну тысячу рублей. Уточняется, что на место происшествия выехала следственная группа, которая установила личность злоумышленника. Им оказался местный житель, ранее судимый за подобное правонарушение.*

Hypothesis: *Ранее местный житель совершал подобное правонарушение.*

Label: *Entailment*

LiDiRus (diagnostic dataset) also belongs to the NLI group of tasks. Additionally, this dataset is annotated with 33 linguistic phenomena under four high-level categories[3]: lexical-semantics, knowledge, logic, and predicate-argument structure. Therefore, LiDiRus can be used to examine the linguistic competence of language models and conduct a systematic analysis of the model behavior. The dataset was translated from English into Russian by professional linguists. During the translation, all the linguistic phenomena in it were preserved.

LiDiRus task example:

Premise: *Кошка сидела на коврике.*

Hypothesis: *Кошка не сидела на коврике.*

Label: *Not entailment*

Logic: *Negation*

[3] Documentation and detailed description of the dataset structure can be found on the project website https://russiansuperglue.com/ru/datasets/

PARus is a binary classification task for accessing commonsense causal reasoning. Each example in the dataset consists of a premise and two alternatives. The task is to select the most probable alternative based on the information from the premise. The correct alternatives are randomized so that the expected accuracy of the random choice guess is 50 percent.

PARus task example:

Premise: *Гости вечеринки прятались за диваном.*

Question: *Что было ПРИЧИНОЙ этого?*

Alternative 1: *Это была вечеринка-сюрприз.*

Alternative 2: *Это был день рождения.*

Correct Alternative: *1*

RUSSE is a binary classification task on a word sense disambiguation problem based on the original RUSSE[4] dataset. Each example contains two sentences and a polysemous word, which occurs in both sentences. The task is determining whether the highlighted word is used in the same sense in both sentences.

RUSSE task example:

Context 1: *Бурые ковровые дорожки заглушали шаги.*

Context 2: *Приятели решили выпить на дорожку в местном баре.*

Word: *дорожка*

Sense match (label): *False*

DaNetQA is a Russian question-answering dataset in the binary classification form. Each example comprises a text fragment and a yes/no question. The task is to answer each question (two possible answers are allowed – YES or NO) based on the information from the given text fragment.

DaNetQA task example:

Text: *В период с 1969 по 1972 год по программе «Аполлон» было выполнено 6 полётов с посадкой на Луне. Всего на Луне высаживались 12 астронавтов США. Список космонавтов Список космонавтов — участников орбитальных космических полётов Список астронавтов США — участников орбитальных космических полётов Список космонавтов СССР и России — участников космических полётов Список женщин-*

[4]https://russe.nlpub.org/downloads/

космонавтов Список космонавтов, посещавших МКС Энциклопедия астронавти.
Question: *Был ли человек на Луне?*
Answer: *Yes.*

MuSeRC is a Machine Reading task. The task is multi-hop in the sense that questions can only be answered using information from multiple sentences. Each example comprises a text, a question, and several answers. For the correct solution, it is necessary to mark all the correct answers.

MuSeRC task example:

Paragraph: *Мужская сборная команда Норвегии по биатлону в рамках этапа Кубка мира в немецком Оберхофе выиграла эстафетную гонку. [...] После этого отставание российской команды от соперников только увеличивалось. Напомним, что днем ранее российские биатлонистки выиграли свою эстафету. В составе сборной России выступали Анна Богалий-Титовец, Анна Булыгина, Ольга Медведцева и Светлана Слепцова. Они опередили своих основных соперниц - немок - всего на 0,3 с*
Question: *На сколько секунд женская команда опередила своих соперниц?*
Candidate answers:

- *Всего на 0,3 секунды. -* ***Label:*** *True*
- *На 0,3 секунды. -* ***Label:*** *True*
- *На секунду. -* ***Label:*** *False*
- *На секунды. -* ***Label:*** *False*

RuCoS is a Machine Reading Comprehension task. Each example consists of a text and cloze-style query with the masked named entity mentioned in the text. The goal is to select from a list of named entities mentioned in the text the one which was initially masked in the query.

RuCoS task example:

Paragraph: *НАСА впервые непосредственно наблюдало «фундаментальный процесс природы». Так специалисты назвали магнитное пересоединение (перестройку силовых линий) полей Солнца и Земли, которое удалось изучить спутникам космического агентства. Посвященное этому исследование опубликовано в журнале Science, кратко о нем сообщает НАСА. Четыре спутника MMS (Magnetospheric Multiscale Mission) совершили в общей сложности более четырех тысяч пролетов через границу магнитосферы планеты. Это позволило ученым непосредственно наблюдать магнитное пересоединение — процесс, в результате которого магнитные линии поля* *сходятся вместе и перестраиваются. Это сопровождается разгоном космических частиц до высоких скоростей.*

Именованные сущности: *НАСА, Солнца, Земли, Science, MMS, Magnetospheric Multiscale Mission*

Query: *В исследовании, опубликованном учеными <placeholder>, изучена динамика этого процесса и показано, что решающий энергетический вклад в физику процесса вносят электроны.*

Correct Entity: *НАСА*

RWSD or Russian Winograd Schema Challenge is devoted to coreference resolution in the binary classification form. The dataset was constructed as a translation of the original Winograd Schema Challenge[5].

RWSD task example:

Text: *Кубок не помещается в чемодан, потому что он слишком большой.*

Span1: *Кубок*

Span2: *он слишком большой*

Corefence (label): *True*

It should be noted that as long as six from nine datasets (RCB, PARus, MuSeRC, TERRa, DaNetQA, RuCoS) in Russian SuperGLUE were not translated from any other language but constructed from Russian sources, RSG takes into account the specifics of the Russian language and, therefore, tests a wide range of NLU aspects that cannot be assessed on translation data only. For example, the tasks mentioned above include texts, related to Russian culture and history; some tasks examples are based on the use of free word order possible in Russian.

4.2.2 Language Model Evaluation on Russian SuperGLUE

To evaluate a language model on the Russian SuperGLUE benchmark, it is necessary to solve all nine benchmark tasks and form predictions for the test set for each task. Each task is evaluated using the corresponding metric (see Table 1). The final result is obtained by averaging the results for all tasks (for the tasks with several metrics, the results of all metrics for this task are preliminarily averaged). In addition, human performance estimates (or human benchmark) were evaluated for all RSG tasks, including the diagnostic dataset LiDiRus. Human

[5] http://commonsensereasoning.org/winograd.html

performance was estimated via *Yandex.Toloka platform*[6]. The total human benchmarks result on RSG is equal to 0.811.

For the convenient use of the benchmark, Russian SuperGLUE platform[7] was developed. It includes the benchmark datasets, and a leaderboard) and proved a user-friendly interface for language model evaluation.

Together with this platform, the *jiant-russian* framework was developed based on [Pruksachatkun Y. et al., 2020] for language model evaluation on the RSG benchmark. This software is implemented as a *Python* library and is available in the project repository. This system allows you to fine-tune Russian and multilingual pre-trained language models from the *HuggingFace library*[8] and evaluate them on Russian SuperGLUE.

4.2.3 Experiments

Table 2. Results of the language model evaluation and human benchmark.

Model	Score	LiDiRus	RCB	PARus	MuSeRC	TERRa	RUSSE	RWSD	DaNetQA	RuCoS
Human Benchmark	0.811	0.626	0.68 / 0.702	0.982	0.806 / 0.42	0.92	0.805	0.84	0.915	0.93 / 0.890
RuBERT (plain)	**0.521**	0.191	0.367 / 0.463	**0.574**	**0.711 / 0.324**	**0.642**	0.726	**0.669**	0.639	**0.32 / 0.314**
RuBERT (conversational)	0.50	0.178	**0.452 / 0.484**	0.508	0.687 / 0.278	0.64	**0.729**	**0.669**	0.606	0.22 / 0.218
mBERT	0.495	**0.189**	0.367 / 0.445	0.528	0.639 / 0.239	0.617	0.69	**0.669**	0.624	0.29 / 0.29
Majority Heuristic	0.468	0.147	0.4 / 0.438	0.478	0.671 / 0.237	0.549	0.595	**0.669**	**0.642**	0.26 / 0.257
TF-IDF	0.434	0.06	0.301 / 0.441	0.486	0.587 / 0.242	0.471	0.57	0.662	0.621	0.26 / 0.252

A series of experiments on language model evaluation on RSG is conducted as part of the thesis research. The following pre-trained language models are tested: RuBERT[9] (plain), RuBERT (conversational)[10], mBERT[11]. The results of the tested models are also compared with the human benchmark), the majority heuristic, and the Naïve Baseline based on TF-IDF text vectorization. All models are evaluated using the RSG methodology (see the previous part of the section). The obtained results are presented in Table 2.

Analysis of the results reveals that at the time of the research, language models perform

[6] https://toloka.yandex.ru/
[7] https://russiansuperglue.com/ru/
[8] https://huggingface.co/models
[9] http://files.deeppavlov.ai/deeppavlov_data/bert/rubert_cased_L-12_H-768_A-12_pt.tar.gz
[10] http://files.deeppavlov.ai/deeppavlov_data/bert/ru_conversational_cased_L-12_H-768_A-12_pt.tar.gz
[11] https://huggingface.co/bert-base-multilingual-cased

significantly worse than the human level. The best result of RuBert (plain) equals 0.521, which is 0.29 lower than the human level (the latter equals 0.811). Nevertheless, models show promising results in the RUSSE and MuSeRC tasks.

In addition, these experiments show that at the time of the research, the tasks presented in the Russian SuperGLUE benchmark can be considered quite complex for language models. That, in turn, positively characterizes the benchmark as a challenge for language models, allowing researchers to evaluate the NLU capabilities of language models at a high level. Moreover, RSG provides an opportunity for an adequate assessment of more advanced language models than those that existed at the time of its creation. Due to the rapid development of NLP in general and language modeling in particular, the latter is highly relevant.

4.3 Language Model Stability Evaluation in Natural Language Inference Task

4.3.1 Problem Statement

This section continues the study of language model evaluation, focusing on language model stability and methods for their stability estimation with respect to different random initializations. Namely, this section is devoted to the research conducted in [Tikhonova M. et al., 2022] on assessing the stability of the BERT language model in *the Natural Language Inference (NLI) task.* As a part of the research, a novel method for the language model stability evaluation is proposed, which makes it possible to estimate how stable a language model learns linguistic features in the process of solving the NLI task. The proposed method is used in a series of multilingual experiments devoted to the study of multilingual BERT[12] (*mBERT*) model stability.

4.3.2 Multilingual Data

In this section, the following multilingual datasets are used.

- **Multilingual diagnostic dataset[13]** in five languages: English, Russian, French, German, and Swedish, which was created specifically for this research. The multilingual diagnostic dataset is based on diagnostic datasets from GLUE and RSG benchmarks (see previous section) for English and Russian, respectively. The latter was additionally translated into the other languages of the study (namely, French, German, and Swedish). The translation was performed by professional linguists with the preservation of the linguistic phenomena marked up in the dataset. Thus, a parallel corpus for the NLI task with linguistic markup in 33 linguistic features was obtained. Such a parallel dataset can be used to evaluate

[12] https://huggingface.co/bert-base-multilingual-cased
[13] https://github.com/MariyaTikhonova/multilingual_diagnostics/

multilingual language models and perform a multilingual comparative analysis in five languages.

- **RTE/TERRa** are the datasets from GLUE и RSG benchmarks, respectively. They are used in the experiments as the primary training data for language model fine-tuning. Similarly to the diagnostic dataset, the TERRa dataset was translated into French, German, and Swedish.
- **MNLI** is a multilingual dataset for the NLI task. It is used as additional training data for model fine-tuning. In the experiments, only English examples from the MNLI dataset (374 thousand English samples) are used.

4.3.3 Metrics

Following the original methodology of GLUE and RSG benchmarks, Matthew's correlation coefficient (MCC) (an analog of metric [Gorodkin J., 2*004*] in the case of binary classification) is used as the main metric for model evaluation. MCC is computed between model predictions and correct answers separately for each of the 33 diagnostic linguistic features.

In addition, for language model stability evaluation with respect to these linguistic features, a stability coefficient (**R**andom **S**eed correlation) is used. This stability coefficient is proposed as a part of the thesis research.

4.3.4 Method for the Stability Evaluation (RScorr Coefficient)

In this research, a method for language model stability evaluation with respect to linguistic features is proposed. It consists of four steps. Below, a short description of the method is given, the pseudocode of the algorithm is presented in Figure 1, and a detailed description of the method can be found in [Tikhonova M. et al., 2022].

<u>Method:</u>

1. Fine-tune a language model *K* time on the same training set with different random initialization[14]:

$$random\_seed = k, k = 0, \dots, K-1.$$

2. For every of *k* runs, the fine-tuned model is evaluated on the diagnostic dataset and a set of MCC coefficients for every linguistic category is computed:

$$MCC_k = (mcc_{1k}, \dots, mcc_{33k}),$$

where mcc_{ik} is MCC for the i^{th} linguistic feature in the k^{th} run.

3. Using the coefficients obtained in step 2, pairwise Pearson correlation coefficients

[14] Random initialization of the additional classification head added for fine-tuning.

between runs are computed:

$$corr_{kj} = PearsonCorr(MCC_k, MCC_j), \forall k, j = 0, \ldots, K-1, k \neq j.$$

4. The final stability correlation coefficient *RScorr* is computed by averaging the correlations from step 3:

$$RScorr = \frac{1}{K(K-1)} \sum_{k \neq j} corr_{kj},$$

Figure 1. Pseudocode of the algorithm for language model stability evaluation with respect to linguistic features in the diagnostic dataset.

```
def compute_stability(pretrained_model, train_data, diagnostics, K):
Input:
pretrained_model is a pre-trained language model
train_data is a training data for model fine-tuning
diagnostics is a diagnostic dataset for model evaluation
K is a number of runs with different random initialization

Output:
RScorr is an average Pearson correlation coefficient between runs with respect to lingustic features

  MCC_coefs = []
  for k in range(K):
    pretrained_model.train(train_data, random_seed = k)
    MCC_k = pretrained_model.evaluate(diagnostics)
    MCC_coefs.append(MCC_k)

  PearsonCorrs = []
  for k in range(K):
    for j in range(K):
      if k != j:
        PearsonCorr_kj = PersonCorrelation(MCC_coefs[k], MCC_coefs[j])

  RScorr = mean(PearsonCorrs)
  return RScorr
```

4.3.5 Influence of the Training Data Volume on Model Stability

The proposed method is used for the mBERT model overall evaluation and to examine the influence of the training data volume on the resulting model stability in five languages. For this, mBERT is fine-tuned in several runs in every language. The model is fine-tuned on the following datasets: RTE/TERRa training data and RTE/TERRa training data augments with the English examples from MNLI. Results are presented in Table 3.

Experiments reveal that the size of the original training data for the NLI task in GLUE and RSG benchmarks is not enough to achieve an adequate level of model stability. Enlarging the training data results in a significant increase in model stability (*RScorr* growth up by 64%) and overall *MCC* (*MCC* improves on average by 49%).

Table 3. Results of the fine-tuning stability w.r.t using additional MNLI training samples in the cross-lingual transfer setting. Overall MCC = overall MCC scores of each model averaged over all model runs. RScorr. = average pairwise Pearson's correlation coefficients between the models' MCC scores in different runs.

Language	Fine-tuning data	Overall MCC	RScorr.
English	**RTE**	0.200 ± 0.016	0.634
	RTE & MNLI	0.294 ± 0.006	0.929
French	**RTE**	0.178 ± 0.027	0.529
	RTE & MNLI	0.268 ± 0.010	0.822
German	**RTE**	0.158 ± 0.024	0.411
	RTE & MNLI	0.213 ± 0.010	0.836
Russian	**RTE**	0.182 ± 0.033	0.455
	RTE & MNLI	0.263 ± 0.012	0.810
Swedish	**RTE**	0.169 ± 0.028	0.517
	RTE & MNLI	0.277 ± 0.016	0.785
Average	**RTE**	0.177 ± 0.017	0.509
	RTE & MNLI	0.236 ± 0.011	0.836

5. Conclusion

The work presents two research projects united by the topic of language model evaluation in Natural Language Understanding. The work is a complete study whose main results are the development of the system for language model evaluation in NLU, a method for language model stability evaluation, and important results related to the stability evaluation of the mBERT model in five languages. The obtained results are approbated in numerous presentations at international scientific conferences, including A-level ones. Their scientific significance is also confirmed by several publications, including two publications with the first authorship of the author of the thesis in Natural Language Engineering (Q1 - Scopus) journal, Proceedings of the International Conference "Dialogue 2021" (Scopus). The developed systems and algorithms are widely used for language model evaluation in various companies and research projects. In particular, more than 2000 different solutions have been submitted to Russian SuperGLUE since its creation. Today, companies like Sber and Yandex use the RSG benchmark to evaluate their models. The theoretical and methodological results obtained in the thesis are used both in Sber and the research studies at HSE University.

As a result of the thesis research, the following **conclusions** can be made.

1. As part of developing the Russian SuperGLUE benchmark for comprehensive language model evaluation, a framework *jiant-russian* for language model evaluation is developed, which can be used to evaluate language models from the HuggingFace project on Russian

SuperGLUE tasks. Jiant-russian makes it possible to fix experimental design and setup and, therefore, ensure the reproducibility of the experiments.

2. With the use of jiant-russian, several pre-trained BERT models are evaluated in Russian. In the experiments, it was shown that in Natural Language Understanding tasks, language models at the time of the research were still far behind the human level (the best model result is 0.521, which is 0.29 lower than the human level of 0.811). Nevertheless, tested models show promising results in word-in-context disambiguation tasks and tasks on machine reading. In addition, such experiment results show that Russian SuperGLUE tasks can be regarded as quite complex, which positively characterizes the benchmark as a challenge for language modeling, allowing researchers to evaluate model NLU capabilities at a high level and providing an opportunity for an adequate assessment of more advanced language models than those that existed at the time of its creation.
3. A novel method for language model stability evaluation is proposed, which makes it possible to evaluate how stable a language model learns linguistic features in the process of solving the NLI task. This result has a methodological significance for language model evaluation.
4. Using the proposed method, a study on the stability of the mBERT linguistic generalization abilities in the NLI task is conducted. A detailed analysis of the impact of the random initialization and the training dataset size is performed. The experiments show that the size of the training data in the standard benchmarks is insufficient for the model stability with respect to learning various linguistic features. However, this stability can be improved for all languages by using additional data only in English. The usage of extra data leads to an overall MCC improvement of 49% and models stability measures via the RScorr coefficient by 64%.

References

[Bowman S. R. et al., 2015] Bowman S. R. et al. A large annotated corpus for learning natural language inference //arXiv preprint arXiv:1508.05326. – 2015.

[Brown T. et al., 2020] Brown T. et al. Language models are few-shot learners //Advances in neural information processing systems. – 2020. – Т. 33. – С. 1877-1901.

[Conneau et al., 2018a] Conneau, Alexis, and Douwe Kiela. "Senteval: An evaluation toolkit for universal sentence representations." arXiv preprint arXiv:1803.05449 (2018).

[Conneau et al., 2018b] Conneau A. et al. XNLI: Evaluating cross-lingual sentence representations //arXiv preprint arXiv:1809.05053. – 2018.

[Cui L. et al., 2020] Cui L. et al. On commonsense cues in BERT for solving commonsense tasks //arXiv preprint arXiv:2008.03945. – 2020.

[Dagan I. et al., 2005] Dagan I., Glickman O., Magnini B. The pascal recognising textual

entailment challenge //Machine learning challenges workshop. – Springer, Berlin, Heidelberg, 2005. – C. 177-190.
[Devlin J. et al., 2019] Devlin J., Chang M.-W., Lee K. and Toutanova K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. 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). Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.
[Dodge J. et al., 2020] Dodge J. et al. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping //arXiv preprint arXiv:2002.06305. – 2020.
[Ettinger A., 2020] Ettinger A. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models //Transactions of the Association for Computational Linguistics. – 2020. – T. 8. – C. 34-48.
[Fenogenova A. et al., 2021] Alena Fenogenova, Tatiana Shavrina, Alexandr Kukushkin, Maria Tikhonova, Anton Emelyanov, Valentin Malykh, Vladislav Mikhailov, Denis Shevelev, Ekaterina Artemova. Russian SuperGLUE 1.1: Revising the Lessons not Learned by Russian NLP-models (2021) A Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2021" Moscow, 2021
[Henderson P. et al., 2018] Henderson P. et al. Deep reinforcement learning that matters //Proceedings of the AAAI conference on artificial intelligence. – 2018. – T. 32. – №. 1.
[Hu J. et al., 2020] Hu J. et al. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation //International Conference on Machine Learning. – PMLR, 2020. – C. 4411-4421.
[Hua H. et al., 2021] Hua H., Li X., Dou D., Xu C. and Luo J. (2021). Noise stability regularization for improving BERT fine-tuning. 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. 3229–3241.
[Kovaleva et al., 2019] Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. 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 4356–4365.
[Goldberg Y., 2019] Goldberg Y. Assessing BERT's syntactic abilities //arXiv preprint arXiv:1901.05287. – 2019.
[Gorodkin J., 2004] Gorodkin J. (2004). Comparing two k-category assignments by a k-category correlation coefficient. Computational Biology and Chemistry 28(5–6), 367–374.
[Konodyuk N. et Tikhonova M., 2022] Konodyuk N., Tikhonova M. Continuous Prompt

Tuning for Russian: How to Learn Prompts Efficiently with RuGPT3? //International Conference on Analysis of Images, Social Networks and Texts. – Springer, Cham, 2022. – C. 30-40.

[Le H. et al., 2019] Le H. et al. Flaubert: Unsupervised language model pre-training for French//arXiv preprint arXiv:1912.05372. – 2019.

[Lee C. et al. 2019] Lee C., Cho K., Kang W. Mixout: Effective regularization to finetune large-scale pretrained language models //arXiv preprint arXiv:1909.11299. – 2019.

[Liang Y. et al., 2020] Liang Y. et al. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation //arXiv preprint arXiv:2004.01401. – 2020.

[Liu X. et al., 2021] Liu X. et al. GPT understands, too //arXiv preprint arXiv:2103.10385. – 2021.

[Madhyastha P. et Jain R., 2019] Madhyastha P., Jain R. On model stability as a function of random seed //arXiv preprint arXiv:1909.10447. – 2019.

[Marelli M. et al., 2014] Marelli M. et al. A SICK cure for the evaluation of compositional distributional semantic models //Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). – 2014. – C. 216-223.

[Pruksachatkun Y. et al., 2020] Pruksachatkun, Yada, et al. "jiant: A software toolkit for research on general-purpose text understanding models." arXiv preprint arXiv:2003.02249 (2020).

[Rogers A. et al., 2020] Rogers A., Kovaleva O., Rumshisky A. A primer in Bertology: What we know about how bert works //Transactions of the Association for Computational Linguistics. – 2020. – T. 8. – C. 842-866.

[Rybak P. et al., 2020] Rybak P. et al. KLEJ: comprehensive benchmark for Polish language understanding //arXiv preprint arXiv:2005.00630. – 2020.

[Shavrina T. et al., 2020] Shavrina T., Fenogenova A., Emelyanov A., Shevelev D., Artemova E., Malykh V., Mikhailov V., Tikhonova M., Chertok A., Evlampiev A. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020. P.

[Storks S. et al., 2019] Storks S., Gao Q., Chai J. Y. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches //arXiv preprint arXiv:1904.01172. – 2019.

[Tikhonova M. et al., 2022] Tikhonova M., Mikhailov, V., Pisarevskaya, D., Malykh, V., Shavrina, T. Ad Astra or astray: Exploring linguistic knowledge of multilingual BERT through NLI task //Natural Language Engineering. – 2022. – C. 1-30.

[Vaswani A. et al., 2017] Vaswani A. et al. Attention is all you need //Advances in neural information processing systems. – 2017. – T. 30.

[Wang, Alex, et al. 2018] Wang, Alex, et al. "GLUE: A multi-task benchmark and analysis platform for natural language understanding." arXiv preprint arXiv:1804.07461 (2018).

[Wang, Alex, et al. 2019] Wang, Alex, et al. "Superglue: A stickier benchmark for general-purpose language understanding systems." Advances in neural information processing systems 32 (2019).

[Warstadt A. et Bowman S., 2019] Warstadt A., Bowman S. R. Linguistic analysis of pretrained sentence encoders with acceptability judgments //arXiv preprint arXiv:1901.03438. – 2019.

[Warstadt et al., 2019] Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, et al. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLPIJCNLP), pages 2870–2880.

[Williams A. et al., 2017] Williams A., Nangia N., Bowman S. R. A broad-coverage challenge corpus for sentence understanding through inference //arXiv preprint arXiv:1704.05426. – 2017.

[Xu L. et al., 2020] Xu L. et al. CLUE: A Chinese language understanding evaluation benchmark //arXiv preprint arXiv:2004.05986. – 2020.