Федеральное государственное автономное образовательное учреждение
высшего образования «Национальный исследовательский университет
«Высшая школа экономики»

На правах рукописи

Тихонова Мария Ивановна

Методы оценивания языковых моделей в задачах понимания естественного языка

РЕЗЮМЕ
диссертации на соискание ученой степени
кандидата компьютерных наук

Москва – 2023

Диссертационная работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Национальный исследовательский университет «Высшая школа экономики»

Научный руководитель:	Воронцов Константин Вячеславович, доктор физико-математических наук, профессор РАН, заместитель заведующего кафедрой «Математические методы прогнозирования» ВМК МГУ, заведующий кафедрой «Машинного обучения и цифровой гуманитаристики» МФТИ, профессор кафедры «Интеллектуальные системы» МФТИ, с.н.с. отдела «Интеллектуальные системы» Вычислительного центра ФИЦ ИУ РАН
Научный руководитель:	Шаврина Татьяна Олеговна, кандидат филологических наук, Руководитель исследовательских проектов, AIRI, Старший научный сотрудник, Институт Языкознания РАН

1. Тема диссертации

Постановка задачи и актуальность исследования

Задача языкового моделирования привлекает большое внимание исследователей последние десятилетия. Языковые модели занимают важную нишу в области *Natural Language Processing* (обработка естественного языка, *NLP*). На сегодняшний день они являются основой решения большого спектра задач, связанных с обработкой естественного языка. К таким задачам относятся всевозможные задачи классификации текстов (задача анализа тональности текста, детекции спама в почте, определение жанра фрагмента текста и другие), задачи, связанные с извлечением информации из текстовых данных (например, разметка по частям речи, извлечение фактов, выделение именованных сущностей и другие), а также широкий спектр задач, связанный с генерацией текста (суммаризация текста, машинный перевод, парафразирование текста и другие).

На сегодняшний день в области NLP преобладает нейросетевой подход к языковому моделированию и существует большое разнообразие нейросетевых языковых моделей.

В связи с данным разнообразием встают вопросы о том, насколько эффективными являются современные модели и насколько хорошо они понимают естественный язык. Таким образом, все большую **актуальность** приобретают вопросы, связанные с оцениванием[1] языковых моделей:

1) возникает необходимость в разработке методов количественного оценивания языковых моделей в различных NLP задачах;
2) возникает необходимость в разработке систем тестов и инструментов, с помощью которой можно оценивать те или иные аспекты языковых моделей и сравнивать их между собой.

Предлагаемое исследование фокусируется на одном из аспектов оценивания языковых моделей. А именно, она посвящена методам оценивания языковых моделей в задачах на понимание естественного языка (*Natural Language Understanding, NLU*).

Цель и задачи диссертационной работы

Основной целью данной работы является разработка методов оценивания языковых моделей в задачах понимания естественного языка и создание необходимого набора инструментов для

[1] здесь и далее термины оценивание и оценка идентичны, так как в литературе распространены оба

осуществления данного оценивания.

Для решения данной цели **ставится следующий ряд задач**:

1. Разработка методов систематического оценивания языковых моделей в задачах понимания естественного языка.
2. Разработка метода оценивания стабильности[2] языковых моделей в задаче распознавания причинно-следственных связей.
3. Проведение серии экспериментов по оценке стабильности поведения мультиязычной языковой модели BERT [Delvin J. et al., 2019] на различных языках. Проверка гипотезы о влиянии объема обучающих данных на стабильность результатов модели и проведение сравнительного анализа между разными языками. Данная серия экспериментов должна быть проведена с использованием метода из пункта 2.

Степень разработанности темы исследования

В данном разделе описаны научные исследования, выполненные к моменту начала диссертационного исследования. Описание степени разработанности разделено по задачам, представленным в разделе «Цель и задачи исследования».

Разработка системы для оценивания языковых моделей в задачах на понимание естественного языка

Основным способом оценивания языковых моделей в задачах на понимание естественного языка на сегодняшний день является подход на основе бенчмарков – наборов из нескольких заданий (тестов), где каждое задание тестирует определенный аспект понимания естественного языка. Для комплексного оценивания на языковая модель должна решить все задания. В последние годы было представлено несколько подобных бенчмарков. SentEval [Conneau et al., 2018a] – один из первых наборов тестов, для оценки качества векторный представлений предложений.

GLUE [Wang et al., 2018] на сегодняшний день является классическим бенчмарком и представляет из себя платформу и набор англоязычных тестов для оценивания языковых моделей на широком спектре задач на понимание естественного языка. Данный подход развивает англоязычный бенчмарк SuperGLUE [Wang, Alex, et al., 2019], включающий более

[2] здесь и далее под стабильностью подразумевается устойчивость модели к изменениям начальной инициализации при дообучении

сложные задачи по сравнению с GLUE. Исследования [Kovaleva et al., 2019; Warstadt et al., 2019] показывают, что GLUE как набор тестов является недостаточно сложным и, как следствие, SuperGLUE является предпочтительным для оценивания языковых моделей.

Существует ряд аналогов GLUE на других языках: FGLUE [Le H. et al., 2019], KLEJ [Rybak P. et al., 2020] и CLUE [Xu L. et al., 2020] – французская, польская и китайская версии бенчмарка, соответственно. А также ряд мультиязычных наборов тестов таких как XGLUE [Liang Y. et al., 2020] и XTREME [Hu J. et al., 2020] для оценивания языковых способностей мультиязычных моделей сразу на нескольких языках.

Однако большинство современных исследований в данной области фокусируются на английском языке и представляют наборы тестов именно для него, в том время как русский язык, затронутый лишь в небольшой части мультиязычных тестов, является недостаточно представленным. На момент начала диссертационного исследования для него не существовало системы тестов для комплексного оценивания способностей языковых моделей на понимание естественного языка, аналогичных GLUE и SuperGLUE для английского языка.

Разработка метода оценивания стабильности языковых моделей в задаче на распознавание причинно-следственных связей и проведение серии экспериментов по оценке стабильности поведения мультиязычной языковой модели BERT

Задаче распознавания причинно-следственных связей *(Natural Language Inference, NLI)* [Storks S. et al., 2019] сегодня уделяется много внимания. Для нее был предложен ряд датасетов, среди которых 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] и XNLI [Conneau et al., 2018b]. Отдельно стоит отметить диагностический датасет, предложенный в рамках GLUE [Wang et al., 2018] бенчмарка, который на сегодняшний день является стандартом для изучения лингвистического знания англоязычных языковых моделей для задачи распознавания причинно-следственных связей.

Стабильность языковых моделей также находится в фокусе внимания современных исследований [Henderson P. et al., 2018; Madhyastha P. et Jain R., 2019; Dodge J. et al., 2020]. В экспериментах [Devlin J. et al., 2019] модель BERT демонстрирует нестабильное поведение при обучении на небольшом объеме данных. Исследования [Lee C. et al., 2019; Mosbach M. et al., 2020; Hua H. et al., 2021] показывают, что изменение случайной инициализации при дообучении модели может вызвать существенные изменения результатов на различных NLP задачах, включая GLUE.

Что касается лингвистического анализа модели BERT и того, как дообучение влияет на знание данной модели, то этому посвящен ряд работ, рассмотренных в обзоре [Rogers A. et al., 2020]. Исследования покрывают различные лингвистические феномены, включая синтаксические свойства [Warstadt A. et Bowman S., 2019], семантическое знание [Goldberg Y., 2019], здравый смысл [Cui L. et al., 2020] и другие [Ettinger A., 2020].

В свете того, что в вышеперечисленных работах отмечается нестабильное, во многом случайное поведение модели BERT, разработка методов оценивания стабильности модели является крайне востребованной, а исследование ее лингвистических способностей является крайне актуальной задачей на сегодняшний день.

Данная научная работа продолжает исследования в данной области, рассматривая стабильность модели BERT в контексте выучивания определенных лингвистических признаков для задачи распознавания причинно-следственных связей.

Научная новизна исследования

1. Впервые предложен метод оценивания стабильности языковых моделей для задачи распознавания причинно-следственных связей.
2. Разработана методология для мультиязычного оценивания моделей на пяти языках с использованием метода из пункта 1.
3. Проведено оригинальное исследование стабильности мультиязычной модели BERT в задаче распознавания причинно-следственных связей на пяти языках и выявлена связь стабильности с размеров набора данных для дообучения модели.
4. В рамках создания первого русскоязычного набора тестов на понимание естественного языка разработан фреймворк для оценивания языковых моделей на данном наборе тестов, с помощью которого проведено оригинальное исследование по оцениванию ряда предобученных моделей архитектуры BERT для русского языка.

2. Основные результаты

Основные положения, выносимые на защиту:

1. В рамках разработки системы русскоязычных тестов *Russian SuperGLUE*[3] *(RSG),* позволяющей производить комплексное оценивание языковых моделей с точки зрения понимания естественного языка, разработан фреймворк *jiant-russian* для оценивания языковых моделей на данном наборе тестов. Данное программное обеспечение позволяет оценивать языковые модели, реализованные на кодовой базе проекта *HuggingFace,* на задачах Russian SuperGLUE. Фреймворк позволяет зафиксировать экспериментальный дизайн оценивания моделей и обеспечить воспроизводимость экспериментов. Тем самым *jiant-russian в* сочетании с Russian SuperGLUE представляет собой удобный инструмент оценивания языковых моделей и их сравнения между собой, что определяет его практическую значимость. С использованием фреймворка проведена серия экспериментов по оцениванию ряда языковых моделей для русского языка. Результаты опубликованы в [Shavrina T. etal., 2020; Fenogenova A. et al., 2021], а фреймворк *jiant-russian* доступен в репозитории [4].
2. Разработан метод оценивания стабильности языковых моделей в задаче распознавания причинно-следственных связей. А именно создан метод, позволяющий оценивать, насколько стабильно языковая модель выучивает различные лингвистические признаки при решении данной задачи. Данный результат имеет теоретическую и методологическую значимость с точки зрения оценивания языковых моделей. Детальное описание метода и полученных результатов опубликовано в [Tikhonova M. et al., 2022].
3. Проведено экспериментальное исследование стабильности мультиязычной языковой модели mBERT на пяти языках в задаче на распознавании причинно-следственных связей. По результатам данной серии экспериментов сделан вывод о том, что базового объема обучающих данных, представленного в классических наборах тестов, недостаточно для того, чтобы данная модель стабильно выучивала лингвистические языковые признаки при решении поставленной задачи. Однако увеличение объема данных для дообучения модели позволяет добиться прироста качества на 49%, и увеличения стабильности результатов (рост

[3] https://russiansuperglue.com/

[4] https://github.com/RussianNLP/RussianSuperGLUE

стабильности на 64%). Детальное описание экспериментов опубликовано в [Tikhonova M. et al. 2022].

Личный вклад в положения, выносимые на защиту

В работе [Tikhonova M. et al., 2022] автором предложен метод оценивания стабильности языковых моделей при решении задачи распознавания причинно-следственных связей. С использованием данного метода основная серия экспериментов по оцениванию стабильности мультиязычной языковой модели BERT на пяти языках и влиянию дополнительных обучающих данных на стабильность модели и ее общий результат.

В рамках проекта по созданию набора тестов Russian SuperGLUE, позволяющих производить комплексную оценку языковых моделей с точки зрения понимания естественного язык, в работе [Shavrina T. et al., 2020] автором разработано ПО (фреймворк) *jiant-russian* для оценивания языковых моделей архитектуры трансформер на данном наборе тестов, с использованием которого, автор проводит серию экспериментов по оцениванию языковых моделей на Russian SuperGLUE. Продолжая данное направление исследований, в рамках работы [Fenogenova A. et al., 2021] автор дорабатывает фреймворк, адаптируя его под изменения, внесенные в набор тестов в рамках данной работы, и добавляет поддержку новых моделей архитектуры трансформер. Помимо этого, автор проводит серию экспериментов по сравнению ряда языковых моделей архитектуры BERT для оценивания их возможностей в задачах на понимание естественного языка.

3. Публикации и апробация работы

Публикации повышенного уровня

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. – С. 1-30. базы данных: **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)**

Публикации стандартного уровня

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 базы данных: 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 базы данных: 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. – С. 30-40. базы данных: Scopus

Доклады на конференциях и семинарах

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

4. Содержание работы

4.1 Постановка задачи языкового моделирования

Объектом данного исследования являются языковые модели. Формально под языковой моделью понимается модель, которая описывает вероятностное распределение на множестве языковых единиц (например, слов, букв, словосочетаний и так далее). Таким образом, языковая модель каждой последовательности языковых единиц $(w_1, \dots, w_n)$ сопоставляет некоторую оценку вероятности этой последовательности $P(w_1, \dots, w_n)$ в языке. В фокусе данного исследования находятся нейросетевые языковые модели на основе архитектуры трансформер [Vaswani A. et al., 2017].

Данная архитектура получила широкое распространение в языковом моделировании и на ее основе предложено множество различных языковых моделей, среди которых такие модели как BERT [Delvin J. et al., 2019], и GPT-3 [Brown T. et al., 2020], рассматриваемые в данном научном исследовании.

4.2 Набор тестов

4.2.1 Задачи Russian SuperGLUE

В рамках диссертационного исследования представлен [Shavrina T. et al., 2020, Fenogenova A. et al., 2021] набор из девяти тестов (бенчмарк) Russian SuperGLUE (RSG) для русского языка, позволяющий проводить комплексную оценку языковой модели с точки зрения понимания естественного языка. Данные задания тестируют различные аспекты понимания естественного языка и могут быть условно разделены на шесть категорий: понимание причинно-следственных связей (*TERRa, RCB),* здравый смысл (*PARus, RUSSE*), знания о мире (*DaNetQA*), машинное чтение (*MuSeRC, RuCos*), логика (*RWSD*) и диагностический датасет *LiDiRus*, дополнительно снабженный лингвистической разметкой на 33 языковых признака. Ниже приведено краткое описание каждого задания, а агрегированная информация о датасетах, их размерах и метриках качества, использованной для оценивания моделей на данном наборе тестов представлена в Таблице 1.

TERRA Задача распознавания причинно-следственных связей в форме бинарной классификации. Каждый пример состоит из двух текстовых фрагментов, для которых необходимо установить наличие/отсутствие причинно-следственной связи.

Таблица 1. Сводные характеристики заданий Russian SuperGLUE. Train/Val/Test - количество примеров в обучающем/валидационном/тестовом наборах данных соответственно. MCC = коэффициент корреляции Мэтьюса, EM = exact match, точное соответствие.

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

Пример задания Terra:

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

Hypothesis: *Автор поста написал про канализацию.*

Label: *Entailment*

RCB Задание на распознавание причинно-следственных связей форме задачи классификации на 3 класса (*enlaiment, contradiction, neutral*).

Пример задания RCB:

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 (диагностический датасет) Также относится к заданиям распознавания причинно-следственных связей. Дополнительно данной набор данных снабжен лингвистической

разметкой[5], включающей 33 языковых признака, разделенные на 4 категории: лексическая семантика (lexical-semantics), знание (knowledge), логика (logic) и предикатно-аргументная структура (predicate-argument structure). Благодаря этому датасет представляет собой удобный инструмент для анализа поведения языковых моделей с точки зрения определенных лингвистических признаков. LiDiRus был переведен с английского языка с сохранением лингвистических признаков, что делает его уникальным инструментом для параллельного мультиязычного анализа.

Пример задания LiDiRus:

Premise: *Кошка сидела на коврике.*

Hypothesis: *Кошка не сидела на коврике.*

Label: *Not entailment*

Logic: *Negation*

PARus Задание бинарной классификации на выбор альтернатив по тексту. Каждый пример содержит посылку, две возможные альтернативы и один из двух типов причинно-следственной связи (причина или следствие). Задача состоит в том, чтобы выбрать альтернативу, которая более вероятно имеет указанную причинно-следственную связь с посылкой.

Пример задания PARus:

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 Задание бинарной классификации на разрешение семантической неоднозначности, созданный на основе RUSSE[6]. Каждый пример содержит 2 предложения и слово, которое употребляется в каждом из них. Необходимо указать, употребляется ли слово в одном значении или разных.

[5] Подробная документация и описание структуры датасета находится на сайте проекта https://russiansuperglue.com/ru/datasets/

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

Пример задания RUSSE:

Context 1: *Бурые ковровые дорожки заглушали шаги.*

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

Word: *дорожка*

Sense match (label): *False*

DaNetQA Русскоязычный вопросно-ответный датасет в форме бинарной классификации. В задании необходимо ответить на закрытый вопрос, подразумевающий бинарный ответ ДА/НЕТ, по текстовому фрагменту.

Пример задания DaNetQA:

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-космонавтов Список космонавтов, посещавших МКС Энциклопедия астронавти.*

Question: *Был ли человек на Луне?*

Answer: *Yes.*

MuSeRC Задание на машинное чтение, в котором по тексту необходимо ответить на вопрос. Каждый пример содержит текст, вопрос и набор вариантов ответов. Для корректного решения необходимо отметить все верные варианты ответов.

Пример задания MuSeRC:

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 Задание на машинное чтение, в котором по тексту необходимо выбрать именованную сущность, о которой идет речь в запросе. Каждое задание состоит из текстового фрагмента, запроса с пропущенной именованой сущностью и списком именованых сущностей, упоминающихся в тексте. Необходимо определить, какая именованая сущность из списка упоминается в запросе.

Пример задания RuCoS:

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 Схема Винограда для русского языка (Russian Winograd Schema Challenge)[7] – аналог теста Тьюринга на машинный интеллект.

Пример задания RWSD:

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

Span1: *Кубок*

Span2: *он слишком большой*

Corefence (label): *True*

Стоит отметить, что благодаря тому, что шесть из девяти датасетов (RCB, PARus, MuSeRC, TERRa, DaNetQA, RuCoS) в данном наборе тестов не являются переводными, а были собраны из русскоязычных источников, RSG во многом учитывает специфику русского языка и тестирует широкий набор аспектов понимания естественного языка, которые невозможно оценить лишь на переводных данных. Например, вышеупомянутые задания содержат тексты, связанные с русской культурой и историей России, а в ряде заданий примеры основаны на использовании свободного порядка слов, допустимого в русском языке.

4.2.2 Оценивание модели на Russian SuperGLUE

Для оценивания языковой модели на RSG необходимо решить все 9 заданий из набора (сформировать предсказания для тестовых наборов данных). После чего каждое задание оценивается с использованием соответствующей метрики (см. Таблицу 1), а итоговый результат получается путем усреднения результатов по всем заданиям (для заданий с несколькими метриками результаты всех метрик для данного задания предварительно усредняются). В дополнении к самому набору тестов, была проведено оценивание уровня человека (human benchmark) с помощью сервиса *Яндекс.Толока*[8], который оказался равным 0.811.

Для удобного использования данного набора тестов была разработана платформа Russian SuperGLUE[9], которая включает наборы данных, таблицу лучших результатов языковых моделей (leaderboard) и предоставляет пользователям удобный интерфейс для оценивания моделей.

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

[8] https://toloka.yandex.ru/

[9] https://russiansuperglue.com/ru/

Вместе с данной платформой был разработан фреймворк *jiant-russian* для оценивания языковых моделей на данном наборе тестов. Данное ПО, основанное на [Pruksachatkun Y. et al., 2020], реализовано в виде библиотеки на языке *Python* и доступно в репозитории проекта. Данная система позволяет дообучать русскоязычные и мультиязычные предобученные языковые модели из библиотеки *HuggingFace*[10].

4.2.3 Эксперименты по оценке моделей на Russian SuperGLUE

В рамках исследования была проведена серия экспериментов по оценке языковых моделей на RSG. Были протестированы следующие предобученные языковые модели: RuBERT[11] (plain), RuBERT (conversational)[12], mBERT[13]. Дополнительно было произведено сравнение результатов данных моделей с результатом человека (human benchmark), выбором большинства (majority heuristic) и методом на основе векторизации текстов с помощью Tf-Idf. Оценивание моделей производилась с использованием методологии RSG (см. предыдущий пункт). Результаты представлены в Таблице 2.

Таблица 2. Результаты оценивания языковых моделей на Russian SuperGLUE и их сравнение с результатами человека.

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

Анализ данных результатов показывает, что языковые модели на момент проведения экспериментов существенно уступают уровню человека в задачах на понимание естественного

[10] https://huggingface.co/models
[11] http://files.deeppavlov.ai/deeppavlov_data/bert/rubert_cased_L-12_H-768_A-12_pt.tar.gz
[12] http://files.deeppavlov.ai/deeppavlov_data/bert/ru_conversational_cased_L-12_H-768_A-12_pt.tar.gz
[13] https://huggingface.co/bert-base-multilingual-cased

языка. Лучший результат RuBert (plain) равен 0.521, что на 0.29 ниже уровня человека, равного 0.811. Тем не менее, модели показывают многообещающие результаты в задании RUSSE на разрешение семантической неоднозначности и задаче на машинное чтение MuSeRC. Помимо этого, данные эксперименты показывают, что на момент проведения исследования задания, представленные в Russian SuperGLUE, являются достаточно сложными с точки зрения языкового моделирования и понимания естественного языка, что в свою очередь положительно характеризует его как сильный бенчмарк, который позволяет оценивать возможности языковых моделей в области понимания естественного языка на высоком уровне и при этом дает возможность для адекватной оценки более продвинутых языковых моделей, чем те, которые существовали на момент его создания. Последнее в силу стремительного развития NLP в общем и языкового моделирования, в частности, является крайне актуальным.

4.3 Оценивание стабильности языковых моделей в задаче распознавания причинно-следственных связей

4.3.1 Постановка задачи

Данный раздел продолжает развитие тему оценивания языковых моделей, фокусируясь на стабильности языковых моделей и методах оценивания их устойчивости относительно различной начальной инициализации. А именно, в данном раздел посвящен оцениванию стабильности языковой модели BERT в задаче распознавания причинно-следственных связей (*Natural Language Inference - NLI*), представленной в работе [Tikhonova M. et al., 2022]. Предложен новый метод оценивания того, насколько стабильно языковые модели выучивают определенные лингвистические признаки при решении задачи NLI, и проведена серия мультиязычных экспериментов по оцениванию мультиязычной модели mBERT[14].

4.3.2 Мультиязычные данные

В данном разделе используются следующие наборы данных:

- **Мультиязычный диагностический датасет[15]** на пяти языках: английском, русском, французском, немецком и шведском. Данный мультиязычный датасет был разработан специально для данного исследования. За основу были взяты существующие диагностические датасеты из бенчмарков GLUE и RSG (см. предыдущий раздел) для

[14] https://huggingface.co/bert-base-multilingual-cased

[15] https://github.com/MariyaTikhonova/multilingual_diagnostics/

английского и русского соответственно. Последний был дополнительно переведен на остальные языки, участвовавшие в данном исследовании, с сохранением лингвистических категорий. Таким образом был получен параллельный корпус для задачи NLI с лингвистической разметкой по 33 языковым категориям, что позволило проводить оценивание модели и производить мультиязычный сравнительный анализ результатов на 5 языках.

- **RTE/TERRa** – набор данных из бенчмарков GLUE и RSG, соответственно, использующиеся в данной серии экспериментов в качестве основного набора данных для дообучения модели. Аналогичного диагностическому датасету, обучающий корпус TERRa был переведен на французский, немецкий и шведский.
- **MNLI** – мультиязычный датасет для задачи NLI. Используется в качестве дополнительного набора данных для дообучения. В эксперименте использовались англоязычные данные, общим объемом 374 тысячи примеров.

4.3.3 Метрики

Аналогично оригинальной методологии GLUE и RSG, для оценки качества был использован коэффициент корреляции Мэтьюса (MCC), аналог R_3 метрики [Gorodkin J., 2*004*] для случая бинарной классификации. MCC вычисляется между предсказаниями модели и истинными ответами для каждой из 33 лингвистических категорий, представленных в диагностике.

Помимо этого, для оценки стабильности поведения модели в отношении лингвистических признаков был использован коэффициент стабильности $RScorr$ (**R**andom **S**eed correlation), предложенный в данном исследовании.

4.3.4 Метод оценивания стабильности (RScorr)

Для того, чтобы оценить стабильность языковой модели в отношении лингвистических признаков был предложен метод, состоящий из четырех шагов. Ниже приведено краткое описание метода, псевдокод представлен на Рисунке 1.

Рисунок 1. Псевдокод алгоритма вычисления стабильности языковой модели на диагностическом датасете в отношении лингвистических категорий.

```
def compute_stability(pretrained_model, train_data, diagnostics, K):
Input:
pretrained_model – предобученная языковая модель
train_data – набор данных для дообучения модели
diagnostics – диагностический датасет для оценки модели
K – число запусков с различной случайной инициализацией

Output:
RScorr – усредненный коэффициент корреляции языковых моделей в отношении языковых категорий для различных запусков

  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
```

Метод:

1. языковая модель дообучается K раз на обучающем наборе данных с различной случайной инициализацией[16]:

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

2. для каждого запуска k производится оценивание модели на диагностическом датасете и вычисляется набор коэффициентов Корреляции Мэтьюса для отдельных лингвистических категорий:

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

где mcc_{ik} – коэффициент корреляции Мэтьюса для категории i при дообучении модели в запуске k

3. с использованием значений с шага 2 вычисляются попарные корреляции Пирсона между различными запусками:

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

[16] Имеется в виду случайная инициализация весов дополнительной классификационной головы модели, которая добавляется при дообучении.

4. итоговый коэффициент стабильности модели получается путем усреднения попарных корреляций с шага 3:

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

4.3.5 Влияние объема данных на стабильность модели

В серии экспериментов предложенный метод был применен для оценивания мультиязычной модели mBERT и влияния объема обучающих данных на стабильность получаемых результатов на пяти языках. Для этого была произведена серия запусков по дообучению модели и ее оценке на всех языках, где в качестве обучающих данных были использованы: данные RTE/TERRa и данные RTE/TERRa, дополненные данными из MNLI. Результаты приведены в Таблице 3.

Таблица 3. Результаты стабильности языковой модели mBERT при дообучении на различном объеме данных в кросс-язычной серии экспериментов. OverallMCC = коэффициент Корреляции Мэтьюса, усредненный по всем запускам. RScorr. = коэффициент стабильности модели в отношении языковых категорий. RTE соответствует обучающим данным из RTE/TERRa в зависимости от языка.

Language	Fine-tuning data	Overall MCC	RS corr.
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

По результатам экспериментов можно сделать вывод о том, что объема обучающих данных, представленных в бенчмарках GLUE и RSG для задачи распознавания причинно-следственных связей недостаточно для достижения адекватного уровня стабильности модели. Добавление дополнительных данных для дообучения позволяет значительно повысить как стабильность модели (рост среднего значения *RScorr* на 64%), так и общий результат модели на диагностики (*MCC* возрастает в среднем на 49%).

5. Выводы

В данной работе выполнено два научных проекта, объединенных общей тематикой, посвященной методам оценивания языковых моделей в задачах распознавания естественного языка. Работа представляет собой законченное исследование, в результате которого была разработана система для оценивания моделей в задачах понимания естественного языка, метод оценивания стабильности языковых моделей и получены важные результаты, связанные с оцениванием стабильности мультиязычной модели Bert на пяти языках. Полученные результаты апробированы на многочисленных выступлениях на международных научных конференциях, в том числе уровня A, и их научная обоснованность подтверждена рядом публикаций, в том числе в двух публикациях с первым авторством соискателя в Natural Language Engineering (Q1 - Scopus), Proceedings of the International Conference “Dialogue 2021” (Scopus). В настоящее время разработанные системы и алгоритмы широко используются для оценивания языковых моделей в различных компаниях и научно-исследовательской деятельности. В частности, с момента создания на Russian SuperGLUE было отправлено более 2000 различных решений, и сегодня его используют для оценивания моделей такие компании как Сбер. Теоретические и методологические результаты, полученные в исследовании, используются как в Сбере, так и в рамках научно-исследовательской деятельности НИУ ВШЭ.
По результатам диссертационного исследования был сделан ряд **выводов:**

1) В рамках разработки системы русскоязычных тестов для комплексного оценивания языковых моделей в задачах понимания естественного языка разработан фреймворк *jiant-russian,* который позволяет оценивать языковые модели, реализованные на кодовой базе проекта *HuggingFace,* на задачах Russian SuperGLUE. Данный фреймворк позволяет зафиксировать экспериментальный дизайн оценивания моделей и обеспечить воспроизводимость экспериментов.

2) С использованием фреймворка *jiant-russian* проведена серия экспериментов по оцениванию ряда предобученных языковых моделей архитектуры BERT для русского языка. В экспериментах было показано, что в задачах на понимание естественного языка данные модели существенно уступают уровню человека (лучший результат среди моделей равен 0.521, что на 0.29 ниже уровня человека, равного 0.811). Тем не менее, они показывают многообещающие результаты в заданиях на разрешение семантической неоднозначности и задачах на машинное чтение. Помимо этого, из экспериментов можно сделать вывод о том, что задания представленные в Russian SuperGLUE являются достаточно сложными для адекватной оценки языковых моделей в области понимания естественного языка, что в свою очередь положительно характеризует его как сильный бенчмарк, который позволяет оценивать возможности языковых моделей в области понимания естественного языка на высоком уровне и при этом дает возможность для адекватной оценки более продвинутых языковых моделей, чем те модели, которые существовали на момент его создания.
3) Предложен метод оценивания стабильности языковых моделей в задаче распознавания причинно-следственных связей, позволяющий оценивать, насколько стабильно языковая модель выучивает различные лингвистические признаки при решении данной задачи.
4) Данный метод был применен для исследования стабильности мультиязычной языковой модели BERT на пяти языках в задаче на распознавание причинно-следственных связей. По результатам данных экспериментов был сделан вывод о том, что стандартного объема обучающих данных, представленного в классических наборах тестов, недостаточно для того, чтобы данная модель стабильно выучивала лингвистические языковые признаки при решении поставленной задачи. Дополнительное исследование данного вопроса показало, что увеличение объема данных для дообучения модели позволяет добиться прироста качества на 49% и значительно увеличения уровня стабильности модели (рост на 64% по метрике *RSCorr*).

Список литературы

[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. – С. 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. – Т. 8. – С. 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. – Т. 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.

[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. – С. 1-30.

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