НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
ноябрь–декабрь 2024 Том 24 № 6 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
November–December 2024 Vol. 24 No 6 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2024-24-6-991-998
УДК 004.89

Параметрический корпус русского языка RuParam

Павел Валерьевич Гращенков[1]✉, Лада Игоревна Паско[2], Ксения Андреевна Студеникина[3], Михаил Михайлович Тихомиров[4]

[1,2,3,4] Московский государственный университет имени М.В. Ломоносова, Москва, 119991, Российская Федерация
[1] Институт востоковедения Российской академии наук, Москва, 107031, Российская Федерация
[1] pavel.gra@gmail.com✉, https://orcid.org/0000-0001-9754-2452
[2] paskolada@yandex.ru, https://orcid.org/0000-0002-0533-809X
[3] xeanst@gmail.com, https://orcid.org/0000-0002-4098-7167
[4] tikhomirov.mm@gmail.com, https://orcid.org/0000-0001-7209-9335

Аннотация

Введение. Основная функция больших языковых моделей заключается в наиболее точной имитации поведения носителей языка. Для того чтобы отслеживать прогресс в решении этой задачи при разработке моделей, а также сравнивать конкурирующие модели между собой, необходимо создание наборов данных для объективной оценки. Распространенный тип таких наборов данных — это корпуса лингвистической приемлемости. Создание таких корпусов основывается на гипотезе о том, что большие языковые модели, как и носители языка, должны быть способны отличать грамматичные предложения от неграмматичных, которые нарушают правила грамматики целевого языка или языков. **Метод.** В работе представлен новый параметрический корпус для русского языка RuParam. Корпус содержит 9,5 тыс. минимальных пар предложений, различающихся по грамматичности, где каждому верному предложению соответствует минимально отличающееся от него ошибочное. Источник неграмматичности в каждой паре сопровождается экспертной лингвистической разметкой. RuParam состоит из двух частей. В первой части используется новый для задачи тестирования больших языковых моделей источник данных — лексико-грамматические тесты по русскому языку как иностранному. Вторая часть состоит из модифицированных корпусных примеров, представляющих грамматические феномены, не входящие в программу преподавания русского языка как иностранного в силу своей сложности. **Основные результаты.** Проведенные эксперименты над моделями показали, что наиболее высокий результат достигается моделями, при обучении которых русскому языку уделялось максимально пристальное внимание на всех этапах обучения, от подготовки данных и токенизации до написания инструкций и обучения с подкреплением (прежде всего YandexGPT и GigaChat). Мультиязычные модели, для которых не было сделано специального акцента на русском языке, показали существенно более низкие результаты. Тем не менее, даже лучшие результаты моделей далеки от оценки людей, которые справляются с задачей практически со 100 % точностью. **Обсуждение.** Ранжирование моделей, полученное в ходе эксперимента, показывает, что разработанный корпус действительно отражает степень владения русским языком. Полученный рейтинг может быть полезен при выборе модели для решения задач обработки естественного языка, где требуется знание грамматики: например, построение морфологических и синтаксических парсеров. В дальнейшем предложенный корпус может быть использован для тестирования собственных моделей.

Ключевые слова

языковые корпуса, русский язык, большие языковые модели, усвоение иностранного языка, обработка естественного языка, оценка приемлемости, универсальная грамматика

Благодарности

Работа выполнена при поддержке Программы развития Московского государственного университета имени М.В. Ломоносова, проект № 23-Ш02-10 «Языковая компетенция носителей естественного языка и нейросетевых моделей». Авторы благодарят студентов Отделения теоретической и прикладной лингвистики МГУ — Марию Кравчук и Даниила Бурмистрова – за существенную помощь в разметке корпуса. Авторы выражают благодарность краудсорсинговой платформе ABC Elementary (https://elementary.center/) за безвозмездное предоставление ресурсов для получения человеческих оценок.

Ссылка для цитирования: Гращенков П.В., Паско Л.И., Студеникина К.А., Тихомиров М.М. Параметрический корпус русского языка RuParam // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24, № 6. С. 991–998. doi: 10.17586/2226-1494-2024-24-6-991-998

Russian parametric corpus RuParam

Pavel V. Grashchenkov[1]✉, Lada I. Pasko[2], Ksenia A. Studenikina[3], Mikhail M. Tikhomirov[4]

[1,2,3,4] Lomonosov Moscow State University, Moscow, 119991, Russian Federation

[1] Institute of Oriental Studies of the Russian Academy of Sciences, Moscow, 107031, Russian Federation

[1] pavel.gra@gmail.com✉, https://orcid.org/0000-0001-9754-2452

[2] paskolada@yandex.ru, https://orcid.org/0000-0002-0533-809X

[3] xeanst@gmail.com, https://orcid.org/0000-0002-4098-7167

[4] tikhomirov.mm@gmail.com, https://orcid.org/0000-0001-7209-9335

Abstract

The main function of large language models is to simulate the behavior of native speakers in the most correct way. Hence, it is necessary to have assessment datasets to track progress in solving this problem as well as regularly compare competing models with each other. There are some datasets of this type, the so-called linguistic acceptability corpora. The hypothesis that underlies the creation of these corpora assumes that large language models, like native speakers, should be able to distinguish correct, grammatical sentences from the ungrammatical ones that violate the grammar of the target language. The paper presents the parametric corpus for Russian, RuParam. Our corpus contains 9.5 thousand minimal pairs of sentences that differ in grammaticality — each correct sentence corresponds to a minimally different erroneous one. The source of ungrammaticality in each pair is supplied with the linguistic markup provided by experts. RuParam consists of two parts: the first part uses a totally new data source for the task of testing large language models — lexical and grammatical tests on Russian as a foreign language. The second part consists of (modified and tagged) examples from real texts that represent grammatical phenomena, not included in the RFL teaching program due to their complexity. As have shown our experiments with different Large Language Models, the highest results are achieved by those models that have been trained on Russian most carefully at all stages, from data preparation and tokenization to writing instructions and reinforcement learning (these are first of all YandexGPT and GigaChat). Multilingual models, which usually receive little or no emphasis on Russian, showed significantly lower results. Still, even the best models results are far from the assessors who completed the task with almost 100 % accuracy. The models ranking obtained during the experiment shows that our corpus reflects actual degree of proficiency in Russian. The resulting rating can be helpful when choosing a model for natural language processing task requiring grammar knowledge: for example, building morphological and syntactic parsers. Thereafter, the proposed corpus can be used to test your own models.

Keywords

corpora, Russian, LLM, L2, natural language processing, acceptability judgements, universal grammar

Acknowledgements

This work was done with the support of MSU Program of Development, Project No. 23-SCH02-10 “Linguistic competence of natural language speakers and neural network models”. We also thank the students of the Department of Theoretical and Applied Linguistics of Lomonosov Moscow State University — Maria Kravchuk and Daniil Burmistrov — for their significant help in markup. We also want to express our gratitude to the ABC Elementary crowdsourcing platform (https://elementary.center/) for the gratuitous provision of resources for obtaining human assessments.

For citation: Grashchenkov P.V., Pasko L.I., Studenikina K.A., Tikhomirov M.M. Russian parametric corpus RuParam. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 6, pp. 991–998 (in Russian). doi: 10.17586/2226-1494-2024-24-6-991-998

Введение

Одно из главных применений больших языковых моделей (БЯМ) состоит в том, чтобы наиболее точно воспроизводить языковое поведение человека. Успешность работы БЯМ обеспечивается, с одной стороны, адекватностью с точки зрения содержания: БЯМ должны владеть верной фактической информацией, быть способными выстраивать причинно-следственные связи и обладать здравым смыслом. С другой стороны, поведение БЯМ должно приближаться к человеческому по формальным лингвистическим критериям: необходимо, чтобы язык моделей был неотличим от человеческого. Носителям языка свойственна способность не только порождать верные высказывания, но и отличать грамматичные предложения от неграмматичных — аналогичные компетенции в идеальном случае ожидаются и от БЯМ. Для того чтобы проверить, насколько качественно задачи, связанные с языковым поведением, выполняются различными моделями, необходимы инструменты объективной оценки. Такие инструменты разрабатываются для каждого из двух аспектов функционирования БЯМ. Семантические и прагматические способности моделей, а также их знания о мире тестируются с помощью таких бенчмарков, как SuperGLUE [1]; для проверки языковых способностей моделей предназначены наборы данных — корпуса лингвистической приемлемости — например, CoLA (Corpus of Linguistic Acceptability) [1] и BLiMP (Benchmark of Linguistic Minimal Pairs) [2]. В настоящей работе предложен новый инструмент для тестирования лингвистической компетенции русскоязычных БЯМ — параметрический корпус минимальных пар RuParam.

Отметим, что применение подобного корпуса не ограничивается проверкой того, насколько хорошо БЯМ владеют естественным языком. Также корпус решает теоретическую задачу параметризации русской грамматики. Так, для того чтобы наше понимание о сильных и слабых сторонах БЯМ было полным, необходимо, чтобы корпус покрывал максимальное число грамматических параметров. Потому проектируемый корпус должен не только содержать необходимые для тестирования эмпирические данные, но и систематизировать грамматику целевого языка.

Исследования лингвистической компетенции языковых моделей. Рассмотрим более подробно существующие корпуса для тестирования лингвистической компетенции моделей. Первый подобный корпус — CoLA [1] — включает в себя предложения на английском языке, взятые из литературы по теоретическому синтаксису. В корпусе представлен широкий круг синтаксических и морфологических феноменов; каждое предложение размечалось экспертами-лингвистами как приемлемое или неприемлемое. Аналогичные корпуса созданы и для других языков, в том числе, для русского — RuCoLA [3]. Хотя корпус CoLA служит отправной точкой при разработке корпусов лингвистической приемлемости, он имеет ряд проблемных мест, которые наследуются его аналогами. Так, предложения берутся из лингвистических работ, где не всегда рассматриваются однозначно приемлемые или неприемлемые языковые явления. Скорее наоборот, интерес лингвистов вызывают сложные неоднозначные феномены, для которых могут возникнуть несогласия между носителями. Таким образом, результаты оказываются «зашумлены» естественной языковой вариативностью.

Другой корпус, который служит эталоном при разработке корпусов приемлемости — BLiMP [2]. В этом корпусе, как и в CoLA, содержатся предложения на английском языке. Однако как источник данных, так и формат получения оценок приемлемости отличается. В настоящей работе предложения были полностью сгенерированы по специальным шаблонам, разработанным лингвистами. Если в CoLA каждое предложение было представлено только одним вариантом — грамматичным или неграмматичным, то в BLiMP содержатся минимальные пары: каждому корректному предложению сопоставляется минимально отличающееся от него предложение с нарушением какого-либо грамматического правила. Это позволяет изолировать вклад конкретной грамматической ошибки от влияния факторов частотности слов и длины предложения. В BLiMP включены 67 невариативных грамматических феноменов. Несмотря на многочисленные достоинства, BLiMP имеет важный недостаток: искусственная генерация примеров делает их менее естественными и надежными для оценки языковых моделей. Аналог BLiMP был совсем недавно разработан и для русского языка — RuBLiMP [4]. В RuBLiMP примеры также создаются по шаблонам, однако в качестве исходных данных используются естественно порожденные предложения из корпусов.

В качестве еще одного источника данных используются корпуса текстов, порожденных людьми, изучающими целевой язык как иностранный. Предложения в таких текстах содержат ошибки и поэтому могут выступать в качестве неприемлемых членов минимальных пар в корпусе. Подобная методика используется в корпусах DaLAJ для шведского [5] и NoCoLA для норвежского [6] языков. Отметим, однако, что в этих корпусах отсутствует грамматическая разметка в строгом смысле: DaLAJ содержит примеры только с лексическими ошибками, а в NoCoLA примеры размечены не по типу грамматической ошибки, а по способу ее исправления.

Таким образом, к недостаткам существующих на данный момент корпусов для оценки лингвистической компетенции моделей относятся следующие проблемы:

— такие корпуса часто включают в себя вариативные параметры, поэтому стандарт сравнения менее однозначен, чем необходимо;
— часто примеры синтезируются искусственно при помощи нейросетей, что ставит под вопрос естественность грамматичных языковых данных;
— в некоторых случаях разметка в корпусах не позволяет исследовать грамматику БЯМ, поскольку в принципе не содержит грамматических категорий, как в DaLAJ или NoCoLA, или содержит чрезвычайно широкие категории, как в RuCoLA, где в качестве тегов выступают, например, «синтаксис» и «семантика»;
— в некоторых корпусах (например, в RuCoLA) наблюдается существенный количественный дисбаланс между грамматичными и неграмматичными примерами;
— во всех тестовых корпусах русского языка грамматическое варьирование достаточно ограниченно, даже для наиболее полного RuBlimp оно исчисляется двенадцатью параметрами.

При создании корпуса RuParam учтены все перечисленные недостатки.

Данные

Общая структура корпуса RuParam. Корпус RuParam располагается в Интернете[1]. Корпус состоит из двух частей, содержащих на данный момент 8248 и 1231 (всего — 9479) примеров, размеченных в общей сложности по 80 макропараметрам. Часть 1 корпуса составляют структурированные данные тестов по русскому языку как иностранному (ТРКИ). Часть 2 получена с помощью модификации предложений из естественных текстов, содержащихся в корпусе RuConst [7]; эта часть корпуса представляет аспекты грамматики, которые занимают важное место в теории языка, но при этом редко встречаются в материалах по РКИ в силу своей сложности. Подобно BLiMP и аналогам, корпус RuParam содержит минимальные пары — каждому грамматичному предложению соответствует неграмматичное. В части 1 корпуса источник неграмматичности в каждой паре определяется в ходе экспертной разметки. В части 2 корпуса неграмматичные примеры

[1] [Электронный ресурс]. Режим доступа: https://github.com/grapaul/RuParam, свободный (дата обращения: 19.11.2024).

создаются лингвистами с учетом заранее определенного набора феноменов. Количество примеров для разных параметров варьирует, но каждый из них представлен не менее чем 10 минимальными парами.

Подкорпус, основанный на данных РКИ. В части 1 корпуса используются материалы лексикограмматического теста ТРКИ, где необходимо выбрать верный способ заполнения пропуска в предложении из нескольких возможных вариантов. Зачастую одному верному предложению соответствует несколько неграмматичных, следовательно, корректное предложение может входить одновременно в несколько минимальных пар. В некоторых случаях нельзя выделить единый источник ошибки, поэтому некоторым парам приписываются сразу несколько категорий (например, «вид» и «время»).

Данная часть корпуса размечена по следующим общим категориям: атрибутивное и предикативное согласование (отдельно с учетом всех грамматических категорий имени); лексическая сочетаемость всех основных частей речи; управление разных частей речи; корректность употребления нефинитных форм; использование вида, времени, модальности; уместное употребление сочинительных и подчинительных союзов; правильность употребления конструкций с числительными; корректность использования той или иной части речи; использование связок в именных предикациях; грамматичность залоговых форм; использование различных типов местоимений (неопределенные, вопросительные и т. д.); грамматичность отрицательных конструкций; другие, более частные категории. В рассматриваемой части корпуса RuParam насчитывается около 50 макропараметров, входящих в эти матричные категории.

Отметим, что в RuParam используется источник примеров, принципиально новый для задачи оценки лингвистической компетенции БЯМ. Данные, связанные с усвоением иностранного языка, и ранее применялись в подобных корпусах (DaLAJ, NoCoLA); однако в этих случаях данные происходили из корпусов ошибок, в качестве исходных выступали некорректные предложения. В корпусе RuParam используются содержащие противопоставление по конкретным лингвистическим параметрам минимальные пары, независимо созданные для задачи, не связанной с решаемой в настоящей работе. Поскольку составители ТРКИ изначально обращают внимание на феномены русского языка, вызывающие трудности у иноязычных учащихся с различным родным языком, эта часть RuParam особенно актуальна при тестировании работы мультиязычных БЯМ.

Приведем некоторые примеры из части 1 корпуса RuParam. Примеры имеют структуру: «id, gram, ungram, label, source», где id — универсальный идентификатор; gram — грамматичный вариант; ungram — неграмматичный вариант; label — обозначание параметра; source — источник примера. Источник для примеров 1–6 — данные РКИ (torfl), в параметр source включено также указание на уровень сложности (A1, B2 и т. д.). Порядок следования примеров в корпусе произвольный, разные грамматические рубрики чередуются, чтобы у БЯМ не накапливалось «догадок» о типе неграмматичности. В публичной версии корпуса, кроме того, грамматичный и неграмматичный примеры даются в случайном порядке, чтобы избежать попадания данных в обучающие выборки моделей.

Пример 1. 1016_1, «Не менее известен Л.Н. Толстой в России и как общественный деятель.», «Не менее известный Л.Н. Толстой в России и как общественный деятель.», форма, torfl_B2.

Пример 2. 957_3, «В прошлом году в стране, решившей развивать пищевую промышленность, был создан специальный комитет в составе правительства.», «В прошлом году в стране, решающей развивать пищевую промышленность, был создан специальный комитет в составе правительства.», вид, torfl_B2.

Пример 3. 963_1, «Собираясь летом путешествовать, я купил всё необходимое весной.», «Собираясь летом путешествовать, всё необходимое куплено весной.», деепр, torfl_B2.

Пример 4. 970_2, «Вчера я встретила подругу, с которой мы не виделись несколько лет.», «Вчера я встретила подругу, которую мы не виделись несколько лет.», упр_глаг, torfl_B2.

1338_2, «В нашей группе много студентов из Кореи.», «В нашей группе много студентов в Корею.», упр_сущ, torfl_A1.

1337_1, «Скажите, пожалуйста, где будет проходить лекция по литературе.», «Скажите, пожалуйста, где будет проходить лекция по литературы.», упр_предл, torfl_A1.

Пример 5. 247_1, «И мы можем интересоваться разными вещами: мой друг любит рисовать, а я играю в футбол.», «И мы могу интересоваться разными вещами: мой друг любит рисовать, а я играю в футбол.», согл_пред_ч, torfl_A1

1327_2, «Почему тебе не нравится эта книга?», «Почему тебе не нравишься эта книга?», согл_пред_л, torfl_A1.

1329_1, «Мне очень нужна твоя помощь.», «Мне очень нужен твоя помощь.», согл_пред_р, torfl_A1.

Пример 6. 1339_1, «Это очень высокое здание.», «Это очень высокая здание.», согл_атр, torfl_A1.

Пример 1 демонстрирует неграмматичность полной формы прилагательного в ряде синтаксических контекстов. Пример 2 связан с необходимостью правильного выбора вида причастий. В примере 3 приводятся верный и неверный случаи употребления подлежащего при деепричастиях. Примеры 4 связаны с выбором правильного управления разных частей речи. Примеры 5 содержат случаи (не)нарушения предикативного согласования по лицу, числу и роду. Пример 6 демонстрирует (не)корректный выбор формы при атрибутивном согласовании.

Подкорпус, основанный на параметрических данных и данных, подчиняющихся универсальным ограничениям. Наличие значительного числа пар предложений, в которых противопоставление осуществляется по наиболее базовым для русского языка параметрам (например, согласование по числу) выгодно отличает RuParam от аналогов, основывающихся на данных из литературы по теоретической лингвистике. Однако не менее важны для оценки компетенций БЯМ

и более сложные лингвистические феномены, которые не преподаются и не тестируются в рамках РКИ.

При создании грамматики некоторого языка количество параметров, которое необходимо проверить и включить в грамматическое описание, исчисляется многими десятками (в проекте wals.info, например, указаны 144 признака, типологически релевантных для языков мира[1]). При разработке корпуса RuParam принят во внимание весь диапазон лингвистических феноменов, встречающихся в корпусах для тестирования БЯМ, и расширен новыми параметрами. Часть примеров отражает универсальные, а не специфичные для русского, синтаксические запреты. К ним относятся островные ограничения [8, 9], запреты, обусловленные принципами связывания [10], универсальное ограничение на непроективные структуры и др.

Представленные в части 2 корпуса данные покрывают следующие области грамматики (которые, в свою очередь, разбиваются примерно на 30 макропараметров, которые, в свою очередь, далее на микропараметры, определяющиеся разными грамматическими категориями, возможностями линейного расположения и т. д.): направление ветвления; связывание; островные ограничения; непроективность; различные виды депиктивов; согласование (предикативное и атрибутивное); употребление разных типов клитик; употребление нефинитных форм; лицензирование единиц с отрицательной полярностью; падеж именной предикации; нулевые подлежащие; залог; вопросительное передвижение. Некоторые из данных категорий встречаются и в части 1 корпуса, однако представлены незначительным числом примеров.

Приведем для части 2 корпуса следующие пары грамматических и неграмматических примеров (структура и другие особенности совпадают с параметрами части 1 корпуса).

Пример 7. 722_1, «Результаты получились самые неожиданные.», «Результаты получились неожиданность.», np_dep, RuConst.

Пример 8. 596_1, «Утвердительный ответ всегда говорит в пользу переизбрания правящей партии.», «Утвердительный ответ всегда говорит в переизбрания правящей партии пользу.», n_gen, RuConst.

732_1, «У нашей сборной было много препятствий, которые мы преодолевали.», «У нашей сборной было много которые мы преодолевали препятствий.», n_rel, RuConst.

810_1, «Переброска осуществляется в соответствии с указом президента Украины.», «Переброска осуществляется указом президента Украины в соответствии с.», p_np, RuConst.

Пример 9. 841_1, «Хотел бы поблагодарить Андрея Шевченко, который помог нам советом.», «Кому хотел бы поблагодарить Андрея Шевченко, который помог советом?», rel_isl, RuConst.

866_1, «Известно, что он родился в 1962 году.», «Когда известно, что он родился?», subj_isl, RuConst.

1013_1, «При этом он затруднился сказать, сколько времени для этого понадобится.», «Для чего при этом он затруднился сказать, сколько времени понадобится?», wh_isl, RuConst.

Пример 10. 896_1, «По его словам, пояс вручил ему тренер Александр Алымов, возглавляющий российское направление единоборства.», «По его словам, пояс вручил ему тренер Александр Алымов, возглавляемый российское направление единоборства.», voice_parta, RuConst.

911_1, «Концовка, придуманная Кристофером Ноланом, вызвала критику ученых, так как такой исход ленты наименее вероятен с научной точки зрения.», «Концовка, придумавшая Кристофером Ноланом, вызвала критику ученых, так как такой исход ленты наименее вероятен с научной точки зрения.», voice_partp, RuConst.

965_1, «Ее использовали, чтобы очищать подъезды от мусора.», «Ее использовались, чтобы очищать подъезды от мусора.», voice_refl, RuConst.

Пример 11. 984_1, «Что мы тогда будем делать?», «Мы тогда будем делать что?», wh, RuConst.

Пример 7 демонстрирует запрет на употребление именных групп в качестве вторичных предикатов в номинативе. Примеры 8 связаны с параметрами направления ветвления в генитивной группе, именной группе с относительным предложением и в предложной группе. Группа примеров 9 демонстрирует некоторые островные ограничения, а именно, остров относительного предложения, подлежащный остров, остров косвенного вопроса. В примерах 10 представлены залоговые пары с активными и пассивными причастиями, а также глаголами на *-ся*. Пример 11 имитирует расположение вопросительного слова на левой периферии или в исходной позиции.

Валидация данных асессорами. Данные (аналогичные использованы далее для эксперимента) были переданы для валидации асессорам, не участвовавшим в подготовке корпуса и не знавшим целей его создания. В оценке принимали участие три эксперта, редакторы по профессии. Примеры подавались экспертам в случайном порядке. Размеченными как неверные оказались 0,6 % ответов от всех экспертов. При допущении, что для каждого примера верным должны быть ответы хотя бы двух из трех асессоров, неверными оказывается 0,3 % примеров (всего 12 (11 из них из РКИ)). Если принять, что все три асессора должны дать верный ответ, оцененными как неверные оказываются 1,5 % примеров (всего 69 (50 из них из РКИ)). 19 примеров части 2 корпуса в этом случае распределены по рубрикам без корреляций. Низкий процент неверных ответов подтверждает корректность разметки корпуса, а наличие ошибок может быть отнесено на счет усталости разметчиков.

Эксперимент по тестированию русскоязычных БЯМ

Дизайн. Для эксперимента были выбраны 7 языковых моделей на корпусе, включающем примерно половину актуальных данных (4689 пар). Экспериментом

[1] Dryer Matthew S. & Haspelmath Martin (eds.) 2013. WALS Online (v2020.3) [Data set]. Zenodo.https://doi.org/10.5281/zenodo.7385533.

было охвачено не все 9479 пар, так как корпус постоянно пополнялся, в том числе уже после проведения первого эксперимента. В тестовый корпус вошли данные по всем основным категориям, представленным в RuParam. Исследование грамматических предпочтений моделей проводилось методом прямого ответа на вербальную инструкцию. Инструкция подавалась модели с каждой парой примеров и предполагала выбор лучшего из двух предложений. Выполнено тестирование нескольких инструкций и в результате была выбрана одна общая, на которой все модели показали наилучший результат в соответствии со следующим примером:

Пример 12. Какое из двух предложений является правильным и грамматичным с точки зрения русского языка?

Предложение 1. {sent_lhs}

Предложение 2. {sent_rhs}

Ответь только одной цифрой 1 или 2, ничего не добавляя.

Все включенные в эксперимент пары примеров были рандомизированы так, чтобы примеры одной категории не шли подряд. Чтобы избежать влияния имеющихся у БЯМ предпочтений к порядку следования примеров, все пары предъявлялись моделям дважды — в обоих вариантах следования приемлемого и неприемлемого предложений.

Насколько авторам настоящей работы известно, такая методика оценки лингвистической компетенции БЯМ применяется впервые. Ранее тестирование моделей либо требовало предварительного дообучения задачам классификации по грамматичности, как в случае CoLA и аналогов, либо предполагало прямое сравнение перплексии членов минимальной пары, как в случае корпусов семейства BLiMP. Разработанная методика не только проверяет способность модели предсказывать, что неграмматичное предложение менее вероятно, чем грамматичное, но и исследует представления моделей о понятиях грамматичности и правильности. По методике происходит обращение к модели так же, как к носителю естественного языка: фактически наша инструкция представляет собой одну из методик экспериментального синтаксиса, *forced choice task*, которая широко применяется при исследовании человеческой грамматики [11]. По этой причине, предложенная методика является наиболее валидной.

В выбранную для эксперимента выборку БЯМ вошли две коммерческие модели, разработанные ведущими российскими IT-компаниями, и 5 моделей, находящихся в открытом доступе. Коммерческие модели — YandexGPT от Яндекса и GigaChat от Сбера, обе принадлежат к семейству GPT3.5 и обучались прежде всего на данных русского языка. На момент тестирования обе модели позиционировались как имеющие 7 млрд параметров. Остальные 5 моделей «видели» русские данные в гораздо меньшем количестве — такая диспропорция должна косвенно подтвердить (или опровергнуть) валидность разработанного корпуса для оценки степени владения русским языком.

Среди 5 моделей свободного распространения были взяты следующие. Первая некоммерческая модель — Openchat-3.5-0106[1] [12] — инструктивная модель на базе Mistral-7B-v0.1, которая показала себя как одна из лучших моделей в данной размерной сетке до выхода LLaMa-3-8B-instruct. Она не обучалось специально на русском языке, однако показывает на нем хорошее качество ответов в сравнении с другими инструктивными версиями на основе Mistral-7B-v0.1. Далее была взята также LLaMa-3-8B-instruct[2] [13] — инструктивная модель на базе LLaMa-3-8B, которая обучалась на рекордных (для открытых моделей) 15T токенах. Сама же инструктивная модель обучалась на корпусе из 10 млн инструкций и на момент выхода была лучшей в своей размерной категории по многим бенчмаркам. Третья некоммерческая модель — Suzume-LLama-3-8B-multilingual[3] [14] — результат дообучения модели LLaMa-3-8B-instruct на 90 тыс. мультиязычных инструкций (включая русский). Еще одна модель — Saiga-LLama3-8b[4] — результат дообучения модели LLaMa-3-8B-instruct на преимущественно русскоязычных инструкциях (датасет saiga_scored). Она показывает более высокое качество по сравнению с LLaMa-3-8B-instruct на многих бенчмарках. Наконец, пятая модель из открытого доступа — Vikhr_5.2[5] [15]. Это единственная модель текущего списка, которая дообучалась на русском языке не только посредством инструкций. Модель основана на Mistral-7B-v0.1 с последующим расширением токенизации токенами русского языка и дообучением на корпусе русскоязычных текстов. Уже после этого модель дообучена на корпусе преимущественно русскоязычных инструкций.

Результаты. Использованные в эксперименте модели справились с поставленной задачей с различной степенью успешности — значения метрики accuracy каждой из моделей приводятся в таблице. Лучший результат показала модель YandexGPT со средним значением 0,927.

Можно отметить следующие закономерности. Во-первых, все модели, кроме GigaChat, лучше справляются с частью 1 корпуса, включающей данные РКИ, чем с частью 2, покрывающей более сложные лингвистические феномены. Эта тенденция особенно хорошо прослеживается для 5 моделей под двойной чертой в таблице. Их средние результаты для каждой из частей корпуса значительно различаются, причем точность на части 2 корпуса часто ненамного отличается от слу-

[1] [Электронный ресурс]. Режим доступа: https://huggingface.co/openchat/openchat-3.5-0106, свободный. Яз. англ. (дата обращения: 19.11.2024).

[2] [Электронный ресурс]. Режим доступа: https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct, свободный. Яз. англ. (дата обращения: 19.11.2024).

[3] [Электронный ресурс]. Режим доступа: https://huggingface.co/lightblue/suzume-llama-3-8B-multilingual, свободный. Яз. англ. (дата обращения: 19.11.2024).

[4] [Электронный ресурс]. Режим доступа: https://huggingface.co/IlyaGusev/saiga_llama3_8b, свободный. Яз. англ. (дата обращения: 19.11.2024).

[5] [Электронный ресурс]. Режим доступа: https://huggingface.co/spaces/Vikhrmodels/small-shlepa-lb, https://huggingface.co/spaces/Vikhrmodels/arenahardlb, свободный. Яз. англ. (дата обращения: 19.11.2024).

Таблица. Результаты тестирования 7 моделей БЯМ на корпусе RuParam
Table. Results of testing seven LLMs on the RuParam corpus

Модель	Весь корпус	Часть 1: РКИ	Часть 2: параметрические и универсальные ограничения		
			Вся часть 2	Параметрические ограничения	Универсальные ограничения
YandexGPT	0,927	0,940	0,888	0,891	0,884
GigaChat	0,895	0,889	0,914	0,906	0,928
Vikhr_5.2	0,758	0,820	0,584	0,587	0,580
Openchat_3.5_0106	0,703	0,754	0,557	0,568	0,540
Saiga_llama3_8b_v7	0,666	0,710	0,542	0,573	0,495
Suzume-llama-3-8B-multilingual	0,637	0,671	0,542	0,553	0,524
LLama3_8b_instr	0,597	0,625	0,518	0,545	0,476

Примечание: голубой цвет соответствует значениям accuracy ниже среднего для данной модели, розовый — выше среднего.

чайного результата (0,5). Во-вторых, можно было бы предположить, что в пределах части 2 корпуса мультиязычные модели будут лучше справляться с подкорпусом универсальных ограничений, чем с подкорпусом типологических параметров, принимающих специфичные для русского языка значения. Однако эта гипотеза не подтверждается — для ряда моделей наблюдается скорее противоположная тенденция. В-третьих, можно утверждать, что чем лучше модель «знакома» с русским языком, тем ближе ее интуиция к интуиции носителей русского языка. YandexGPT и GigaChat существенно превосходят остальные модели, прежде всего — Openchat_3.5_0106, Saiga_LLama3_8b_v7, Suzume-LLama-3-8B-multilingual, LLama3_8b_instr. Модель Vikhr_5.2, обученная с русскоязычной токенизацией на русских текстах и инструкциях, располагается между YandexGPT/GigaChat и четверкой мультиязычных моделей. Такая иерархизация позволяет говорить о том, что разработанный корпус действительно отражает степень владения русским языком. Наконец, даже лучшие результаты существенно отстают от близкой к единице оценки асессоров-людей, что также представляется важным результатом. Как бы хорошо ни были обучены БЯМ на данных целевого языка, их «языковая интуиция» на данный момент все-таки отличается от таковой у людей.

Заключение

В работе представлен RuParam — параметрический корпус для тестирования компетенций больших языковых моделей в области владения русским языком. Разработанный корпус RuParam справляется с рядом проблем, свойственных некоторым предшественникам. Корпус включает в основном однозначные, а не вариативные лингвистические параметры; вошедшие в корпус данные были порождены носителями языка независимо — либо экспертами при разработке тестов по русскому языку как иностранному, либо в ходе естественной языковой деятельности; корпус снабжен обширной и подробной грамматической разметкой, разработанной на основе лингвистической теории и типологии; благодаря формату минимальных пар грамматичные и неграмматичные предложения содержатся в корпусе в равном количестве и могут быть сопоставлены напрямую. В части 1 корпуса использован новый для задачи тестирования лингвистической компетенции больших языковых моделей источник данных — тесты для иноязычных учащихся. Часть 2 RuParam обеспечивает включение в корпус широкого круга грамматических феноменов: как специфичных для русского языка, так и универсальных.

Выполнен эксперимент с 7 большими языковыми моделями. С одной стороны, он показал, что RuParam пригоден для тестирования компетенций больших языковых моделей в области русского языка: модели решают задачу с разным качеством, однако ни одна модель не достигает уровня носителя языка. С другой стороны, в ходе эксперимента получена объективная оценка владения русским языком для ряда больших моделей. В эксперименте использована ранее не применявшаяся методика тестирования, которая опирается на текстовый формат предъявления инструкции и моделирует извлечения данных о грамматической компетенции у носителей-людей.

В дальнейшем развитие RuParam продолжится. В частности, планируется дополнить список грамматических параметров, количественно расширить корпус новыми примерами, рассмотреть зависимость успешности решения задачи моделями от уровня тестов владения русским языком как иностранным, а также провести более подробное (по отдельным параметрам) исследование индивидуальных грамматик конкретных моделей.

Литература

1. Warstadt A., Singh A., Bowman S.R. Neural network acceptability judgments // Transactions of the Association for Computational Linguistics. 2019. V. 7. P. 625–641. https://doi.org/10.1162/tacl_a_00290
2. Warstadt A., Parrish A., Liu H., Mohananey A., Peng W., Wang S.-F., Bowman S.R. BLiMP: The benchmark of linguistic minimal pairs for English // Transactions of the Association for Computational Linguistics. 2020. V. 8. P. 377–392. https://doi.org/10.1162/tacl_a_00321
3. Mikhailov V., Shamardina T., Ryabinin M., Pestova A., Smurov I., Artemova E. RuCoLA: Russian Corpus of Linguistic Acceptability // Proc. of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022. P. 5207–5227. https://doi.org/10.18653/v1/2022.emnlp-main.348
4. Taktasheva E., Bazhukov M., Koncha K., Fenogenova A., Artemova E., Mikhailov V. RuBLiMP: Russian benchmark of linguistic minimal pairs // Proc. of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024. P. 9268–9299.
5. Volodina E., Mohammed Y.A., Klezl J. DaLAJ — a dataset for linguistic acceptability judgments for Swedish // Proc. of the 10th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2021). 2021. P. 28–37.
6. Jentoft M., Samuel D. NoCoLA: The Norwegian corpus of linguistic acceptability // Proc. of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa). 2023. P. 610–617.
7. Гращенков П.В. RuConst: Синтаксический корпус русского с разметкой по непосредственным составляющим // Вестник Московского университета. Серия 9. Филология. 2024. № 3. С. 94–112. https://doi.org/10.55959/MSU0130-0075-9-2024-47-03-7
8. Ross J.R. Constraints on variables in syntax: PhD thesis / Massachusetts Institute of Technology. 1967. 523 p.
9. Белова Д.Д., Вознесенская А.Ю., Герасимова А.А. и др. Русские острова в свете экспериментальных данных. М.: Буки Веди, 2021. 412 с.
10. Chomsky N. Lectures on Government and Binding: The Pisa Lectures. Dordrecht: Walter de Gruyter GmbH & Company KG, 1981. 371 p.
11. Schütze C., Sprouse J. Judgment data // Research Methods in Linguistics. Cambridge: Cambridge University Press, 2014. P. 27–50.
12. Wang G., Cheng S., Zhan X., Li X., Song S., Liu Y. OpenChat: Advancing open-source language models with mixed-quality data // arXiv. 2023. arXiv:2309.11235. https://doi.org/10.48550/arXiv.2309.11235
13. Grattafiori A., Dubey A., Jauhri A., Pandey A. et al. The Llama 3 Herd of Models // ArXiv. 2024. ArXiv:2407.21783. https://doi.org/10.48550/arXiv.2407.21783
14. Devine P. Tagengo: A multilingual chat dataset // Proc. of the Fourth Workshop on Multilingual Representation Learning (MRL 2024). 2024. P. 106–113.
15. Nikolich A., Korolev K., Shelmanov A. Vikhr: The family of open-source instruction-tuned large language models for Russian // arXiv. 2024. arXiv:2405.13929v2. https://doi.org/10.48550/arXiv.2405.13929

Авторы

Гращенков Павел Валерьевич — доктор филологических наук, доцент, Московский государственный университет имени М.В. Ломоносова, Москва, 119991, Российская Федерация; ведущий научный сотрудник, Институт востоковедения Российской академии наук, Москва, 107031, Российская Федерация, SC 54683976400, https://orcid.org/0000-0001-9754-2452, pavel.gra@gmail.com

Паско Лада Игоревна — студент, Московский государственный университет имени М.В. Ломоносова, Москва, 119991, Российская Федерация, https://orcid.org/0000-0002-0533-809X, SC 58871267300, paskolada@yandex.ru

Студеникина Ксения Андреевна — программист, Московский государственный университет имени М.В. Ломоносова, Москва, 119991, Российская Федерация, SC 57424115900, https://orcid.org/0000-0002-4098-7167, xeanst@gmail.com

Тихомиров Михаил Михайлович — кандидат физико-математических наук, научный сотрудник, Московский государственный университет имени М.В. Ломоносова, Москва, 119991, Российская Федерация, SC 55699714800, https://orcid.org/0000-0001-7209-9335, tikhomirov.mm@gmail.com

Статья поступила в редакцию 18.09.2024
Одобрена после рецензирования 10.10.2024
Принята к печати 25.11.2024

References

1. Warstadt A., Singh A., Bowman S.R. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 2019, vol. 7, pp. 625–641. https://doi.org/10.1162/tacl_a_00290
2. Warstadt A., Parrish A., Liu H., Mohananey A., Peng W., Wang S.-F., Bowman S.R. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 2020, vol. 8, pp. 377–392. https://doi.org/10.1162/tacl_a_00321
3. Mikhailov V., Shamardina T., Ryabinin M., Pestova A., Smurov I., Artemova E. RuCoLA: Russian Corpus of Linguistic Acceptability. *Proc. of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 5207–5227. https://doi.org/10.18653/v1/2022.emnlp-main.348
4. Taktasheva E., Bazhukov M., Koncha K., Fenogenova A., Artemova E., Mikhailov V. RuBLiMP: Russian benchmark of linguistic minimal pairs. *Proc. of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 9268–9299.
5. Volodina E., Mohammed Y.A., Klezl J. DaLAJ — a dataset for linguistic acceptability judgments for Swedish. *Proc. of the 10th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2021)*, 2021, pp. 28–37.
6. Jentoft M., Samuel D. NoCoLA: The Norwegian corpus of linguistic acceptability. *Proc. of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa),* 2023, pp. 610–617.
7. Grashchenkov P.V. RuConst: A Treebank for Russian. *Lomonosov Philology Journal. Series 9. Philology*, 2024, vol. 3, pp. 94–112. (in Russian). https://doi.org/10.55959/MSU0130-0075-9-2024-47-03-7
8. Ross J.R. *Constraints on variables in syntax*: PhD thesis. Massachusetts Institute of Technology. 1967, 523 p.
9. Belova D.D., Voznesenskaia A.Iu., Gerasimova A.A. et al. *Russian Islands in the Light of the Experimental Data*. Moscow, Buki Vedi Publ., 2021, 412 p. (in Russian)
10. Chomsky N. *Lectures on Government and Binding*: The Pisa Lectures. Dordrecht, Walter de Gruyter GmbH & Company KG, 1981, 371 p.
11. Schütze C., Sprouse J. Judgment data. *Research Methods in Linguistics*. Cambridge, Cambridge University Press, 2014, pp. 27–50.
12. Wang G., Cheng S., Zhan X., Li X., Song S., Liu Y. OpenChat: Advancing open-source language models with mixed-quality data. *arXiv*, 2023, arXiv:2309.11235. https://doi.org/10.48550/arXiv.2309.11235
13. Grattafiori A., Dubey A., Jauhri A., Pandey A. et al. The Llama 3 Herd of Models. *arXiv*, 2024, arXiv:2407.21783. https://doi.org/10.48550/arXiv.2407.21783
14. Devine P. Tagengo: A multilingual chat dataset. *Proc. of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, 2024, pp. 106–113.
15. Nikolich A., Korolev K., Shelmanov A. Vikhr: The family of open-source instruction-tuned large language models for Russian. *arXiv*, 2024, arXiv:2405.13929v2. https://doi.org/10.48550/arXiv.2405.13929

Authors

Pavel V. Grashchenkov — D.Sc. (Philology), Associate Professor, Lomonosov Moscow State University, Moscow, 119991, Russian Federation; Leading researcher, Institute of Oriental Studies of the Russian Academy of Sciences, Moscow, 107031, Russian Federation, SC 54683976400, https://orcid.org/0000-0001-9754-2452, pavel.gra@gmail.com

Lada I. Pasko — Student, Lomonosov Moscow State University, Moscow, 119991, Russian Federation, https://orcid.org/0000-0002-0533-809X, SC 58871267300, paskolada@yandex.ru

Ksenia A. Studenikina — Software Developer, Lomonosov Moscow State University, Moscow, 119991, Russian Federation, SC 57424115900, https://orcid.org/0000-0002-4098-7167, xeanst@gmail.com

Mikhail M. Tikhomirov — PhD (Physics & Mathematics), Scientific Researcher, Lomonosov Moscow State University, Moscow, 119991, Russian Federation, SC 55699714800, https://orcid.org/0000-0001-7209-9335, tikhomirov.mm@gmail.com

Received 18.09.2024
Approved after reviewing 10.10.2024
Accepted 25.11.2024