ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

На правах рукописи

Михайлов Владислав Николаевич

ЭТАЛОННОЕ ТЕСТИРОВАНИЕ ЯЗЫКОВЫХ МОДЕЛЕЙ НА 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: Артемова Екатерина Леонидовна, кандидат технических наук, «Национальный исследовательский университет «Высшая школа экономики», LMU Munich.

1. Тема диссертации

Обработка естественного языка (англ. *Natural Language Processing, NLP*) является междисциплинарным направлением математической лингвистики, компьютерных наук и искусственного интеллекта, ориентированным на разработку технологий для выполнения задач, которые требуют знание и понимание языка, например: машинный перевод, извлечение информации, распознавание грамматических ошибок и суммаризация текста [30]. Большие языковые модели на основе архитектуры Трансформер [115] (далее «языковые модели») стали неотъемлемой частью решений таких задач и привели к смене парадигмы в области NLP. Такие модели предобучены на больших корпусах текстов и эффективно адаптируются для решения целевых задач с помощью методов дообучения или обобщения с использованием нескольких примеров-демонстраций. Опережающее развитие языковых моделей обусловливает необходимость разработки стандартных методов объективной оценки их обобщающей способности в зависимости от типа задачи, предметной области и языка.

Эталонное тестирование, под которым понимается бенчмаркинг (англ. *benchmarking)*, получило широкое признание в области компьютерных наук с 1960-х годов как традиционный подход к сравнению технологий по различным критериям, таким как качество решения задачи, вычислительная эффективность, безопасность и стабильность [54; 59]. Эталонное тестирование осуществляется с помощью т.н. бенчмарков, которые представляют собой коллекцию стандартных наборов данных, метрик качества и метода агрегации результатов. Сообщество NLP разработало более 2000 бенчмарков[1], оказавших значительное влияние на развитие языковых технологий и охватывающих различные аспекты их оценки, включая, но не ограничиваясь общим пониманием естественного языка [117; 119], языковой компетенцией [123], переносом знания между языками [7], устойчивостью к состязательным атакам [120], вычислительной эффективностью [141] и предубеждениями относительно социальных меньшинств [72]. Держатели большинства NLP бенчмарков предоставляют открытые рейтинги систем, которые позволяют проводить сравнительную оценку языковых моделей относительно друг друга и качества выполнения задачи людьми. Несмотря на то что бенчмарки стали более проблемно-ориентированным [61; 89], они разработаны для малого количества языков [49] и используют методы агрегации результатов, не соответствующие многообразию критериев оценки [26].

Данное диссертационное исследование посвящено эталонному тестированию языковых моделей на задачах понимания естественного языка. Мы предлагаем три NLU бенчмарка, впервые разработанных на материале русского языка. Бенчмарки включают в себя задачи машинного чтения, разрешения лексической многозначности, разрешения кореференции, определения логической и причинно-следственной связи, оценки приемлемости предложения и распознавания сгенерированного текста. Последняя задача представляет особый интерес для области генерации естественного языка (англ. *natural language generation, NLG*) в связи с возрастающими

[1] paperswithcode.com/area/natural-language-processing.

рисками использования генеративных языковых моделей в злонамеренных целях [126]. Вместе с бенчмарком для оценки распознавания сгенерированного текста мы представляем новый подход к решению этой задачи, в основе которого лежат методы топологического анализа данных (англ. *topological data analysis, TDA*; [19]). Кроме того, в данном диссертационном исследовании предложен фреймворк для агрегации результатов эталонного тестирования с помощью принципов теории многокритериального принятия решений [6]. Разработанные методы агрегации результатов более устойчивы, чем общепринятый метод агрегации на основе среднего арифметического, и могут быть использованы для ранжирования языковых моделей в разнообразных экспериментальных постановках.

Актуальность

Бенчмарки на материале русского языка. Совершенствование технологий машинного обучения (МО) неразрывно связано с надежной методологией оценки. Область NLP в основном сосредоточена на английском языке и имеет неравномерное распределение наборов данных среди более 7000 языков [13; 88]. Исследовательское направление переноса знания между языками ориентировано на решение проблемы недостатка данных с помощью мультиязычных языковых моделей. Один из самых стандартных подходов заключается в дообучении таких моделей на обучающей выборке высокоресурсного языка – чаще всего английского – и оценке на тестовой выборке другого языка [90]. Несмотря на то что данный подход имеет большой потенциал, он обладает несколькими недостатками. Качество переноса знания зависит от степени близости исходного и целевого языков, а так же от объема текстовых данных на целевом языке в предобучающем корпусе языковой модели [58]. Вместе с тем языки, типологически близкие к английскому, хорошо представлены в мультиязычных бенчмарках, таких как XGLUE [62] и XTREME [45; 91], в то время как другие языки ограничены малым количеством задач в связи с отсутствием качественных наборов данных.

В последние годы методология эталонного тестирования на материале английского языка была адаптирована для многих типологически разных языков, таких как польский [92], корейский [78], баскский [114], арабский [99], словенский [135], китайский [131], японский [55], персидский [52], и индонезийский [128]. Однако русский язык является одним из тех языков, которые получили наименьшее внимание в отношении разработки стандартных ресурсов оценки языковых моделей [49]. В данном диссертационном исследовании впервые предложены три NLU бенчмарка на материале русского языка.

1. Russian SuperGLUE [103] — это коллекция девяти наборов данных для оценки понимания русского языка, созданных с нуля и разработанных по аналогии с англоязычным бенчмарком SuperGLUE [117]. Предложенный бенчмарк включает в себя задачи машинного чтения, разрешения лексической многозначности, разрешения кореференции, определения логической и причинно-следственной связи, а так же диагностический набор данных для лингвистической интерпретации языковых моделей. Результаты эмпрической оценки на мо-

мент публикации показывают, что языковые модели значительно уступают людям в решении задач NLU. В течение трех последних лет было разработано большое количество языковых моделей для русского языка, которые могут превосходить уровень человека на определенных задачах, но все еще отстают от него в среднем на 4,9%.

2. RuCoLA (Russian Corpus of Linguistic Acceptability; [68]) — это корпус предложений на русском языке, размеченных по бинарной шкале приемлемости. Приемлемость предложения определяется его грамматичностью и осмысленностью с точки зрения носителя языка [21]. RuCoLA состоит из внутридоменных предложений, собранных вручную из публикаций и методических материалов по фундаментальной лингвистике, и внедоменных предложений, сгенерированных девятью нейросетевыми моделями для русского языка. Набор внедоменных предложений создан для прикладного применения оценки приемлемости предложения в задачах генерации русского языка. Проведенная эмпирическая оценка показывает, что широко используемые языковые модели заметно уступают людям, особенно в обнаружении морфологических и семантических ошибок, и перенос знания грамматики между русским, английским [124] и итальянским [109] языками едва возможен. Однако наблюдается незначительная разница между качеством на внутридоменных и внедоменных наборах предложений, что свидетельствует о том, что большинство языковых моделей хорошо обобщаются при оценке приемлемости сгенерированных предложений.
3. RuATD (Russian Artificial Text Detection; [101]) — это бенчмарк, состоящий из сгенерированных текстов и естественных текстов, собранных из открытых источников, относящихся к разным предметным областям. Сгенерированные тексты получены с помощью 13 нейросетевых моделей для русского языка, дообученных на задачах суммаризации текста, генерации парафраза, упрощения текста и машинного перевода. Кроме того, для генерации текстов используются подходы обратного перевода и открытой генерации. Бенчмарк RuATD предложен в рамках кампании Dialogue Evaluation в 2022 году как открытое соревнование, которое включает в себя две задачи: (i) распознавание сгенерированного текста и (ii) определение автора текста. В общей сложности, соревнование привлекло внимание 38 решений: 30 для первой постановки задачи и 8 – для второй. Разница в качестве между лучшими и худшими решениями в обеих задачах составила примерно 20% правильно предсказанных ответов (далее «accuracy »). Результаты оценки решений позволяют сделать вывод о том, что людям сложно распознать сгенерированный текст, в то время как модели достигают 83% accuracy.

Распознавание сгенерированных текстов. Предупреждение: Текст, выделенный коричневым цветом сгенерирован моделью ChatGPT[2], чтобы продемонстрировать необходимость разработки устойчивых детекторов сгенерированного текста. Языковые модели стали мощным инструментом для генерации текста, но их злонамеренное использование может привести к

[2] openai.com/blog/chatgpt

серьезным последствиям, например, к распространению предубеждений относительно социальных меньшинств, присутствующих в предобучающем корпусе, генерации ложной информации и нарушению конфиденциальности пользовательских данных. Поэтому важно разрабатывать и использовать языковые модели ответственно, тщательно оценивая потенциальные риски.

Развитие языковых моделей способствует появлению новых стратегий их злонамеренного использования и вместе с тем мотивирует разработку инновационных подходов к снижению потенциальных рисков [15]. В рамках этого исследовательского направления мы предлагаем новый детектор сгенерированного текста, в основе которого лежат методы TDA [56]. Детектор представляет собой линейный классификатор, обученный на признаках TDA, извлеченных из графового представления карт внимания языковых моделей. Признаки TDA включают в себя стандартные свойства графов, описательные характеристики баркодов, а так же признаки на основе расстояния до паттернов внимания (англ. *attention patterns*; [24]). Экспериментальные результаты показывают, что предложенный детектор превосходит базовые решения на основе счетных векторных представлений и языковой модели BERT [32] на 10% accuracy на наборах данных из трех предметных областей (Reddit, отзывы на продукты Amazon и новостные статьи). Кроме того, детектор на основе TDA лучше обобщается при распознавании текстов, сгенерированных моделями GPT-2, которые отсутствовали в обучающей выборке классификаторов [84]. Диагностическое тестирование признаков TDA показывает их чувствительность к поверхностным и синтаксическим признакам, которая более подробно анализируется в последующей научно-исследовательской работе [20].

Методы агрегации результатов. Вопрос о том, является ли среднее аримфетическое подходящим методом агрегации результатов эталонного тестирования, остается открытым. Среднее арифметическое упрощает эталонное тестирование языковых моделей, несмотря на значительные усилия сообщества NLP разрабатывать бенчмарки, соответствующие текущему уровню развития области. В частности, данный метод агрегации не учитывает критерии оценки, такие как сложность задачи и предметная область [69; 125]. Более того, системы NLP могут превосходить друг друга по качеству только на конкретных задачах, что приводит к смещенной оценке при усреднении результатов [1; 73].

Применяя устоявшиеся практики из теории многокритериального принятия решений, мы предлагаем фреймворк Vote'n'Rank [85], который может использоваться для ранжирования языковых моделей на бенчмарках, состоящих из множества задач, а так же при эталонном тестировании, учитывающем множество критериев оценки. Vote'n'Rank включает в себя восемь методов агрегации, которые учитывают ранжирование систем по каждому критерию оценки и подходят для агрегации неоднородных метрик качества. Эмпирическая оценка предложенных методов агрегации проведена в сравнении с методами на основе среднего арифметического и среднего геометрического в четырех экспериментальных постановках: (i) переранжирование открытых рейтингов систем GLUE, SuperGLUE и VALUE [60], (ii) определение условий, обеспе-

чивающих первое место системы в рейтинге, (iii) оценка устойчивости методов агрегации к пропущенным значениям метрик качества и (iv) ранжирование языковых моделей с учетом предпочтений пользователя. Методы агрегации Vote'n'Rank более устойчивы, чем базовые решения, и обеспечивают интерпретируемость получаемых результатов.

Цель исследования. Цель данного диссертационного исследования заключается в разработке стандартных ресурсов и инструментов для эталонного тестирования, которые в долгосрочной перспективе обеспечат (i) исчерпывающее сравнение существующих и новых языковых моделей для русского языка относительно уровня человека, (ii) репрезентативность русского языка в исследовательских направлениях по переносу знания между языками и (iii) прикладное применение эталонного тестирования, распознавания сгенерированного текста и оценки генерации естественного языка.

2. Основные результаты и выводы

Основной **вклад** данного диссертационного исследования заключается в следующем:

1. Предложены бенчмарки Russian SuperGLUE, RuCoLA и RuATD для оценки обобщающей способности языковых моделей на 11 разнообразных задачах понимания русского языка. Кроме того, разработаны методология оценки людей на данных задачах и методология сбора и разметки текстовых данных с учетом особенностей русского языка. Каждый бенчмарк имеет открытый рейтинг систем для определения достигнутого прогресса в развитии языковых моделей.
2. Вместе с бенчмарком RuATD, разработан новый метод автоматического распознавания сгенерированного текста, который опирается на извлечение геометрических и структурных свойств карт внимания языковых моделей с помощью методов TDA.
3. Предложен фреймворк Vote'n'Rank, который включает в себя восемь методов агрегации на основе теории многокритериального принятия решений для ранжирования языковых моделей при проведении эталонного тестирования. Помимо этого, предоставлены практические рекомендации по использованию фреймворка, которые опираются на теоретические свойства методов агрегации результатов и сценарии потенциального применения.
4. С использованием предложенных бенчмарков и инструментов эталонного тестирования проведен подробный сравнительный анализ более 100 систем NLP относительно уровня человека в различных экспериментальных постановках, включая модели на основе счетных векторных представлений, моноязычные и мультиязычные языковые модели, их ансамбли и другие конфигурации.

Теоретическая и практическая значимость. Мы вносим проблемно-ориентированный вклад в направление эталонного тестирования языковых моделей на основе теоретических концепций из области лингвистики, TDA и теории многокритериального принятия решений. Следующие факторы определяют *значимость* данного диссертационного исследования. Предложенные бенчмарки, кодовая база, открытые рейтинги систем, проекты по оценке людей на задачах NLU и другие материалы находятся в открытом доступе под лицензией Apache 2.0:

- Бенчмарк Russian SuperGLUE (GitHub; russiansuperglue.com)
- Бенчмарк RuCoLA (GitHub; rucola-benchmark.com)
- Бенчмарк RuATD (GitHub)
    1. Распознавание сгенерированного текста: kaggle.com/competitions/ruatd-binary
    2. Определение автора текста: kaggle.com/competitions/ruatd-authorship
- Детектор сгенерированного текста (GitHub)
- Фреймворк Vote'n'Rank (GitHub)

Предложенные бенчмарки стали стандартными ресурсами для эталонного тестирования предобученных моделей для русского языка, на которых было оценено более 2000 закрытых решений от академического сообщества и индустриальных компаний. В общей сложности на открытых рейтингах систем представлено более 90 систем NLP, отранжированных по качеству решения задач NLU относительно уровня человека, включая широко используемые языковые модели, например: RuLeanALBERT[3], ruGPT-3[4], YaLM[5], FRED-T5[6], и ruRoBERTa[7]. Проекты по оценке людей на задачах NLU могут быть переиспользованы для многих исследовательских целей, таких как воспроизводимость результатов оценки людей [12], анализ влияния дизайна проекта на целевое качество [79] и изучение отличий в достигаемом качестве на задачах NLU, выполняемыми разметчиками без предметной компетенции и разметчиками-экспертами [50].

С помощью фреймворка Vote'n'Rank исследователи и разработчики могут проводить эталонное тестирование систем независимо от области МО. Фреймворк позволяет использовать собственные данные и определять предпочтения в ранжировании систем. Предложенные инструменты и ресурсы для эталонного тестирования также могут использоваться в образовательных целях, например, для практики разработки и оценки моделей машинного и глубокого обучения.

[3] hf.co/yandex/RuLeanALBERT

[4] hf.co/ai-forever/rugpt3large_based_on_gpt2

[5] hf.co/yandex/yalm-100b

[6] hf.co/ai-forever/FRED-T5-1.7B

[7] hf.co/ai-forever/ruRoBERTa-large

Наконец, что не менее важно, RuCoLA и RuATD способствуют развитию моделей для оценки грамматичности и осмысленности сгенерированных текстов на русском языке (например, ruRoBERTa-large-rucola[8]), распознавания пропаганды, распространяемой с помощью генеративных языковых моделей, и предупреждения пользователей о потенциально сгенерированном медиа-контенте в социальных сетях и на новостных платформах.

Результаты, выносимые на защиту.

1. Russian SuperGLUE, RuCoLA, и RuATD — ресурсы для эталонного тестирования языковых моделей на материале русского языка.
2. Интерпретируемый и устойчивый метод для распознавания сгенерированного текста на основе TDA.
3. Фреймворк Vote'n'Rank для ранжирования и определения ведущих систем NLP на бенчмарках, состоящих из множества задач.
4. Эмпирическая оценка более 100 языковых моделей и их конфигураций на задачах NLU.

Личный вклад в результаты, выносимые на защиту. Данная диссертация включает в себя шесть научно-исследовательских публикаций, которые являются результатом междисциплинарного сотрудничества между экспертами в различных областях. В первой публикации [103], автором диссертации разработан RuCoS (англ. *Russian Reading Comprehension with Commonsense*) — самый большой набор данных на материале русского языка для задачи машинного чтения, который включен в бенчмарк Russian SuperGLUE. Вторая публикация [37] подробно описывает разработанные автором методологии создания RuCoS и оценки людей на данной задаче. Кроме того, автор провел эмипирическую оценку базовых решений на наборе данных RuCoS.

Вклад автора в третьей публикации [68] заключается в разработке дизайна бенчмарка RuCoLA, сборе внутридоменных предложений из корпуса тестовых заданий единого государственного экзамена по русскому языку [102] и методических материалов для проекта корпусного описания русской грамматики, разработке методологий разметки внедоменных предложений и оценки людей на данной задаче (совместно с Татьяной Шамардиной), определении статистических и лингвистических критериев контроля качества данных, а так же проведении анализа ошибок и анализа результатов эмпирической оценки языковых моделей совместно с Максимом Рябининым.

В четвертой публикации [101] автор разработал дизайн бенчмарка RuATD, внес вклад в качестве соруководителя проекта, агрегировал текстовые данные, собранные соавторами, и разработал методологию оценки людей на задаче распознавания сгенерированного текста совместно с Татьяной Шамардиной и Аленой Феногеновой. Вклад автора в пятой публикации [56] заключается в разработке экспериментального дизайна, проведении экспериментов по диа-

[8] hf.co/RussianNLP/ruRoBERTa-large-rucola

гностическому тестированию предложенных признаков TDA и анализе результатов каждого эксперимента, описанного в работе.

В шестой публикации [85] автор внес вклад как соруководитель проекта, разработал экспериментальный дизайн и провел эксперименты по переранжированию бенчмарков при содействии Марка Рофина. Более того, автор внес значительный вклад в написание текста каждой публикации.

Публикации и апробация работы

* означает равный вклад соавторов

Публикации повышенного уровня

1. *Татьяна Шаврина, Алена Феногенова, Антон Емельянов, Денис Шевелев, Екатерина Артемова, Валентин Малых, Владислав Михайлов, Мария Тихонова, Андрей Черток и Андрей Евлампиев.* 2020. Russian SuperGLUE: Бенчмарк для оценки общего понимания русского языка (Russian SuperGLUE: A Russian Language Understanding Evaluation Benchmark). В материалах конференции по эмпирическим методам в области обработки естественного языка 2020 года (EMNLP), стр. 4717—4726. Онлайн. Ассоциация по компьютерной лингвистике. Конференция ранга A по рейтингу CORE.
2. *Алена Феногенова, Владислав Михайлов и Денис Шевелев.* 2020. MuSeRC and RuCoS: Наборы данных для задачи машинного чтения на материале русского языка (Read and Reason with MuSeRC and RuCoS: Datasets for Machine Reading Comprehension for Russian). В материалах 28-й международной конференции по компьютерной лингвистике (COLING), стр. 6481—6497. Онлайн и Барселона, Испания. Международный комитет по компьютерной лингвистике. Конференция ранга A по рейтингу CORE.
3. *Владислав Михайлов*, Татьяна Шамардина*, Максим Рябинин*, Алена Пестова, Иван Смуров и Екатерина Артемова.* 2022. RuCoLA: Бенчмарк для оценки приемлемости предложения (RuCoLA: Russian Corpus of Linguistic Acceptability). В материалах конференции по эмпирическим методам в области обработки естественного языка 2022 года (EMNLP), стр. 5207—5227. Онлайн и Абу-Даби, Объединенные Арабские Эмираты. Ассоциация по компьютерной лингвистике. Конференция ранга A по рейтингу CORE.
4. *Лаида Кушнарева*, Даниил Чернявский*, Владислав Михайлов*, Екатерина Артемова, Сергей Баранников, Александр Бернштейн, Ирина Пионтковская, Дмитрий Пионтковский и Евгений Бурнаев.* 2021. Распознавание сгенерированного текста с помощью топологического анализа карт внимания (Artificial Text Detection via Examining the Topology of Attention Maps). В материалах конференции по эмпирическим методам в области обработки естественного языка 2021 года (EMNLP), стр. 635—649. Онлайн и Пунта-Кана, Доминиканская республика. Ассоциация по компьютерной лингвистике. Конференция ранга A по рейтингу CORE.

5. *Марк Рофин*, Владислав Михайлов*, Михаил Флоринский*, Андрей Кравченко, Елена Тутубалина, Татьяна Шаврина, Даниел Карабекян и Екатерина Артемова.* 2023. Vote'n'Rank: Пересмотр методологии эталонного тестирования с помощью теории многокритериального принятия решений (Vote'n'Rank: Revision of Benchmarking with Social Choice Theory). В материалах 17-ой европейской конференции по компьютерной лингвистике (EACL). Онлайн и Дубровник, Хорватиия. Ассоциация по компьютерной лингвистике. Конференция ранга А по рейтингу CORE.

Публикации стандартного уровня

1. *Татьяна Шамардина*, Владислав Михайлов*, Даниил Чернявский, Алена Феногенова, Марат Саидов, Анастасия Валеева, Татьяна Шаврина, Иван Смуров, Елена Тутубалина и Екатерина Артемова.* 2022. RuATD: Соревнование по автоматическому распознаванию сгенерированных текстов (Findings of the RuATD Shared Task 2022 on Artificial Text Detection in Russian). В материалах международной конференции по компьютерной лингвистике и интеллектуальным технологиям "Диалог 2022". Проиндексирована в Scopus.

Доклады на научных конференциях и семинарах

1. Russian SuperGLUE: Бенчмарк для оценки общего понимания русского языка. Семинар в научно-учебной лаборатории моделей и методов вычислительной прагматики, НИУ ВШЭ.
2. Russian SuperGLUE: Бенчмарк для оценки общего понимания русского языка. EMNLP. 17 ноября 2020. Онлайн-доклад.
3. Все способы измерить слона: Russian SuperGLUE и RuSentEval. Симпозиум по анализу больших данных для выявления глобальных вызовов и трендов в сфере человеческого потенциала. Институт статистических исследований и экономики знаний НИУ ВШЭ. 12 апреля 2021, Москва, Россия.
4. MuSeRC and RuCoS: Наборы данных для задачи машинного чтения на материале русского языка. COLING. 8 декабря 2020. Онлайн-доклад.
5. Russian Commitment Bank: Уроки машинного обучения против уроков лингвистики — все не выучено? Moscow HSE Pragmatics Workshop. 30 сентября 2021. Онлайн-доклад.
6. Распознавание сгенерированного текста с помощью топологического анализа карт внимания. EMNLP. Онлайн и Пунта-Кана, Доминиканская республика, 7 ноября 2021. Устный доклад.
7. RuATD-2022: Соревнование по автоматическому распознаванию сгенерированных текстов. Конференция "Диалог 2022". 16 июня 2022. Онлайн-доклад.
8. RuCoLA: Бенчмарк для оценки приемлемости предложения. EMNLP. Онлайн и Абу-Даби, Объединенные Арабские Эмираты, 9 декабря 2022. Стендовый доклад.

9. Vote'n'Rank: Пересмотр методологии эталонного тестирования с помощью теории многокритериального принятия решений. EACL. 2 мая 2023. Онлайн-доклад.

Автор так же является соорганизатором научных и обучающих семинаров, смежных теме данного диссертационного исследования

1. *Татьяна Шаврина, Владислав Михайлов, Валентин Малых, Екатерина Артемова, Олег Сериков и Виталий Протасов.* 2022. NLP Power! Первый научный семинар по эффективному эталонному тестированию в области NLP (NLP Power! The First Workshop on Efficient Benchmarking in NLP). Ассоциация по компьютерной лингвистике (ACL), Дублин, Ирландия. Конференция ранга A* по рейтингу CORE.
2. *Адаку Ученду, Владислав Михайлов, Чжуен Ли, Саранья Венкатраман, Татьяна Шаврина и Екатерина Артемова.* 2022. Обучающий семинар по распознаванию сгенерированного текста (Tutorial on Artificial Text Detection). 15-ая международная конференция по генерации естественного языка (INLG). Онлайн и Уотэрвилл, Мэн, США. Ассоциация по компьютерной лингвистике. Конференция ранга B по рейтингу CORE.

Объем и структура работы. Диссертация содержит введение, содержание публикаций и заключение. Полный объем диссертации составляет 158 страниц.

3. Содержание работы

3.1. Russian SuperGLUE: Бенчмарк для оценки общего понимания русского языка

Набор данных	|Train|	|Dev|	|Test|	Задача	Метрики	Предметная область
DaNetQA	392	295	295	MRC	Acc.	Википедия
MuSeRC	500	100	322	MRC	$F1_a$/EM	новостные статьи, сказки, академ. тексты, худ. лит.,
RuCoS	72k	4.3k	4.1k	MRC	F1/EM	новостные статьи (Лента.ру, Deutsche Welle)
RUSSE	19.8k	8.5k	12.1k	WSD	Acc.	Википедия, НКРЯ, словарные статьи
PARus	400	100	500	NLI	Acc.	блоги, энциклопедические тексты
RWSD	606	204	154	Coref.	Acc.	худ. лит.
RCB	438	220	348	NLI	F1/Acc.	новостные статьи, худ. лит.
TERRa	2616	307	3198	NLI	Acc.	новостные статьи, худ. лит.
LiDiRus	✗	✗	1104	NLI	MCC	новостные статьи, Википедия, Reddit, академ. тексты

Таблица 1: Задачи, представленные в Russian SuperGLUE. LiDiRus — это диагностический набор данных. **Обозначения**: MRC=машинное чтение; WSD=разрешение лексической многозначности; НКРЯ=Национальный корпус русского языка; Coref.=разрешение кореференции; NLI=определение логической связи. **Метрики**: F1=F1-score; $F1_a$=macro-average F1 [51]; Acc.=доля правильных ответов; EM=полное совпадение; MCC=Коэффициент корреляции Мэтьюса [65].

Мотивация создания бенчмарка Russian SuperGLUE аналогична GLUE [119] и SuperGLUE [117]: разработать стандартную методологию для измерения развития технологий понимания русского языка. Russian SuperGLUE (см. Таб. 1) — это набор из восьми задач NLU, охватывающих

различные предметные области, размеры наборов данных и формулировки задач. Мы предоставляем веб-сайт и открытый рейтинг систем для оценки, сравнения и анализа языковых моделей на закрытых тестовых выборках.

3.1.1. Метод

DaNetQA (A Yes/No Вопрос Answering Dataset) — это задача машинного чтения, сформулированная как задача бинарной классификации, в которой модель должна ответить "да" (`True`) или "нет" (**False**) на вопрос по тексту из статьи Википедии.

- **Текст**: "*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.*"
- **Вопрос**: "*Был ли человек на луне?*"
- **Ответ**: `True`

MuSeRC (Russian Multi-Sentence Reading Comprehension) — это задача машинного чтения, решение которой требует способностей к агрегации информации из нескольких предложений и логическому рассуждению.

- **Text**: "*(1) Мужская сборная команда Норвегии по биатлону в рамках этапа Кубка мира в немецком Оберхофе выиграла эстафетную гонку. (2) Вторыми стали французы, а бронзу получила немецкая команда. (3) Российские биатлонисты не смогли побороться даже за четвертое место, отстав от норвежцев более чем на две минуты. <...> (11) Напомним, что днем ранее российские биатлонистки выиграли свою эстафету. (12) В составе сборной России выступали Анна Богалий-Титовец, Анна Булыгина, Ольга Медведцева и Светлана Слепцова. (13) Они опередили своих основных соперниц - немок - всего на 0,3 секунды.*"
- **Вопрос**: "*На сколько секунд женская команда опередила своих соперниц?*"
- **Варианты ответа**:
  - ☑ "*Всего на 0,3 секунды.*"
  - ☑ "*На 0,3 секунды.*"
  - ☐ "*На секунду.*"
  - ☐ "*На 0.5 секунд.*"

RuCoS (Russian Reading Comprehension with Commonsense Reasoning) — это задача машинного чтения, решение которой требует знания о мире. Каждый пример состоит из отрывка из новостной статьи, предложения с пропуском и вариантов ответа. Задача заключается в том, чтобы заполнить пропуск в предложении, выбрав один или несколько подходящих вариантов ответа, встретившихся в тексте.

- **Текст**: "*Мать двух мальчиков, брошенных отцом в московском аэропорту Шереметьево, забрала их. Об этом сообщили ТАСС в пресс-службе министерства образования и науки*

Хабаровского края. <...> Также министерство социальной защиты населения рассматривает вопрос о бесплатном оздоровлении детей в летнее время. Через несколько дней после того, как Виктор Гаврилов бросил своих детей в аэропорту, он явился с повинной к следователям в городе Батайске Ростовской области.

  - *Бросившего детей в Шереметьево отца задержали за насилие над женой*
  - *Россиянина заподозрили в истязании брошенных в Шереметьево детей*
  - *Оставивший двоих детей в Шереметьево россиянин сам пришел к следователям*"

- **Предложение**: "*26 января ___ бросил сыновей в возрасте пяти и семи лет в Шереметьево.*"
- **Ответ**: "*Виктор Гаврилов*"

RUSSE (Russian Words in Context) — это задача разрешения лексической многозначности, в которой от модели требуется определить, используется ли данное слово с одним и тем же значением в паре предложений.

- **Предложение 1**: "*Бурые ковровые дорожки заглушали шаги.*"
- **Предложение 2**: "*Приятели решили выпить на дорожку в местном баре.*"
- **Слово**: "*дорожка*"
- **Ответ**: `False`

PARus (Choice of Plausible Alternatives for Russian) — это задача определения причинно-следственных связи в формате бинарной классификации. Модель получает на вход предложение и два варианта ответа и должна определить, какой из вариантов ответа является причиной, а какое — следствием ситуации, описанной в предложении.

- **Предложение**: "*Гости вечеринки прятались за диваном.*"
- **Вариант ответа 1**: "*Это была вечеринка-сюрприз.*"
- **Вариант ответ 2**: "*Это был день рождения.*"
- **Тип связи**: "*Причина*"
- **Ответ**: "*Это была вечеринка-сюрприз.*"

RWSD (Russian Winograd Schema Challenge) — это задача разрешения кореференции, в которой модели на вход подается предложение, референт и местоимение или именная группа. Модель должна определить, есть ли между ними кореферентная связь.

- **Text**: "*Кубок не помещается в коричневый чемодан, потому что он слишком большой.*"
- **Местоимение/именная группа**: "*он слишком большой*"
- **Референт**: "*чемодан*"
- **Ответ**: `False`

Категория	Феномены
Лексическая семантика	Кванторы, именованные сущности, лексическое следование, симметрия, фактивность, морфологическое отрицание, избыточность
Предикатно-аргументная структура	совпадение и несовпадение ролей ключевых аргументов глагола, предложные группы, интерсективность и неинтерсективность, рестриктивность, анафора и кореферентность, согласование, активный и пассивный залог, эллипсис, номинализация, относительная клауза, дативные конструкции, генитив и партитив
Формальная семантика	отрицание и двойное отрицание, интервалы и числа, восходящая и нисходящая монотонность, немонотонность, различие в глагольном времени, конъюнкция и дизъюнкция, условные конструкции, универсальные и экзистенциальные предложения
Знания	здравый смысл, знания о мире

Таблица 2: Лингвистические феномены, представленные в LiDiRus.

RCB (Russian Commitment Bank) — это задача определения логической и причинно-следственной связи в формате многоклассовой классификации. Каждый пример состоит из предпосылки и гипотезы. Модель должна предсказать, подтверждается ли гипотеза, противоречит ли она предпосылке или является нейтральной по отношению к ней.

- **Предпосылка**: "*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.*"
- **Гипотеза**: "*Ранее местный житель совершал подобное правонарушение.*"
- **Ответ**: `entailment`

TERRa (Textual Entailment Recognition for Russian) — это задача бинарной классификации, в которой требуется определить, может ли значение одного предложения быть выведено из другого в паре предложений.

- **Предложение 1**: "*Автор поста написал в комментарии, что прорвалась канализация.*"
- **Предложение 2**: "*Автор поста написал про канализацию.*"
- **Ответ**: `entailment`

LiDiRus (Linguistic Diagnostic for Russian) — это диагностический набор данных, который покрывает широкий спектр лингвистических феноменов для интерпретации модели (см. Таб. 2). LiDiRuS позволяет оценить взаимосвязь между предсказаниями модели и наличием во входных примерах феноменов с помощью анализа корреляции.

- **Предложение 1**: "*Мы построили наше общество на неэкологичной энергии.*"
- **Предложение 2**: "*Мы построили наше общество на экологичной энергии.*"
- **Ответ**: `not entailment`
- **Лексическая семантика**: *морфологическое отрицание*
- **Формальная семантика**: *отрицание*

Модель	Общий рез-т	LiDiRus MCC	RCB F1/Acc.	PARus Acc.	MuSeRC $F1_a$/EM	TERRa Acc.	RUSSE Acc.	RWSD Acc.	DaNetQA Acc.	RuCoS F1/EM
TF-IDF	43.4	5.9	30.1/44.1	48.6	58.7/24.2	47.1	66.0	66.2	62.1	25.6/25.1
ruBERT	54.6	18.6	43.2/46.8	61.0	65.6/25.6	63.9	**89.4**	67.5	74.9	25.5/25.1
mBERT	54.2	15.7	38.3/42.9	58.8	62.6/25.3	62.0	84.0	67.5	79.0	37.1/36.7
Человек	**80.2**	**62.6**	**68.0/70.2**	**98.2**	**80.6/42.0**	**92.0**	74.7	**84.0**	**87.9**	**93.0/92.4**

Таблица 3: Результаты эмпирической оценки базовых решений на закрытых тестовых выборках Russian SuperGLUE и диагностическом наборе данных. **Метрики**: F1=F1-мера; $F1_a$=макро-усредненная F1-мера [51]; Acc.=доля правильных ответов; EM=полное совпадение; MCC=Коэффициент корреляции Мэтьюса [65]. Значения метрик умножены на 100. Значения, выделенные жирным начертанием, означают лучший результат на наборе данных.

3.1.2. Эмпирическая оценка

Базовые решения. Мы проводим эмпирическую оценку базовых решений на основе счетных векторных представлений и языковой модели BERT для русского языка. TF-IDF представляет собой логистическую регрессию, обученную на признаках TF-IDF [94], посчитанных на подвыборке из 20 тысяч русскоязычных и англоязычных статей из Википедии. mBERT — это мультиязычная языковая модель BERT, предобученная на моноязычных корпусах Википедии на 104 языках. ruBERT-base [134] — это языковая модель BERT, предобученная на корпусах новостных текстов и Википедии для русского языка. Модели BERT независимо дообучены на каждой задаче. Мы так же проводим оценку решения задач людьми с помощью Толоки[9], платформы для разметки данных. Инструкции по разметке и примеры веб-интерфейса доступны в репозитории GitHub[10].

Ключевые результаты. Результаты эмпирической оценки представлены в Таб. 3. Языковые модели BERT значительно уступают по качеству людям на большинстве предложенных задач. Однако языковые модели решают задачу RUSSE лучше людей на 14.7% accuracy. Сравнивая результаты между языковыми моделями, мы находим, что ruBERT-base решает задачи немного лучше, чем mBERT, в особенности задачи машинного чтения (RuCoS, DaNetQA) и задачи определения логической и причинно-следственной связи (RCB, TERRa).

3.1.3. Ретроспектива

С момента выхода публикации Russian SuperGLUE прошел валидацию сообществом с последующим улучшением наборов данных [38]: (i) расширение тестовой выборки RUSSE 6700 на примеров и увеличение качества решения задачи людьми на новой закрытой тестовой выборке на 6%, (ii) расширение набора данных MuSeRC на 300 примеров, (iii) увеличение валидационной

[9] toloka.ai

[10] github.com/RussianNLP/RussianSuperGLUE/HumanBenchmark

Ранк	Модель	Общий рез-т	LiDiRus MCC	RCB F1/Acc.	PARus Acc.	MuSeRC $F1_a$/EM	TERRa Acc.	RUSSE Acc.	RWSD Acc.	DaNetQA Acc.	RuCoS F1/EM
1	Человек	**81.1**	**62.6**	**68.0**/**70.2**	**98.2**	80.6/42.0	**92.0**	80.5	**84.0**	**91.5**	**93.0**/89.0
2	FRED-T5 1.7B FT	76.2	49.7	49.7/54.1	84.2	91.6/77.3	87.1	**82.3**	66.9	88.9	90.0/90.2
3	Golden Transformer v2.0	75.5	51.5	38.4/53.4	90.6	**93.6**/**80.4**	87.7	68.7	64.3	91.1	92.0/**92.4**
4	YaLM p-tune	71.1	36.4	35.7/47.9	83.4	89.2/70.7	84.1	71.0	66.9	85.0	92.0/91.6
5	FRED-T5 large FT	70.6	38.9	45.6/54.6	77.6	88.7/67.8	80.1	77.5	66.9	79.9	87.0/86.3
6	RuLeanALBERT	69.8	40.3	36.1/41.3	79.6	87.4/65.4	81.2	78.9	66.9	76.0	90.0/90.2
7	FRED-T5 1.7B encoder FT	69.4	42.1	31.1/44.1	80.6	88.2/66.6	83.1	72.3	66.9	73.5	91.0/91.1
8	ruT5-large FT	68.6	32.0	45.0/53.2	76.4	85.5/60.8	77.5	77.3	66.9	79.0	86.0/85.9
9	ruRoBERTa-large FT	68.4	34.3	35.7/51.8	72.2	86.1/63.0	80.1	74.8	66.9	82.0	87.0/86.7

Таблица 4: Топ-девять позиций на открытом рейтинге систем Russian SuperGLUE. **Метрики**: F1=F1-мера; $F1_a$=макро-усредненная F1-мера [51]; Acc.=доля правильных ответов; EM=полное совпадение; MCC=Коэффициент корреляции Мэтьюса [65]. Под FT подразумевается дообучение. Значения метрик умножены на 100. Значения, выделенные жирным начертанием, означают лучший результат наа наборе данных.

и закрытой тестовой выборок RuCoS в два раза, (iv) увеличение размера обучающей, валидационной и закрытой тестовой выборок DaNetQA до 1750, 821 и 805 примеров соответственно, и (v) исправление неточностей, а также улучшение консистентности разметки путем ручной валидации наборов данных MuSeRC и RuCoS.

За последние три года разница в качестве решения задач между людьми и языковыми моделями сократилась с 25,8 до 4,9. Прогресс был достигнут благодаря стремительному развитию области и разработке новых предобученных языковых моделей для русского языка, таких как RuLeanALBERT, ruGPT-3, YaLM, FRED-T5 и ruRoBERTa. На бенчмарке Russian SuperGLUE было оценено более 2000 закрытых решений от академического сообщества и индустриальных компаний. На открытом рейтинге представлено 28 систем NLP (см. Таб. 4), включая энкодеры (ruRoBERTa-large, RuLeanALBERT) энкодеры-декодеры (ruT5-large, FRED-T5) и ансамбли (Golden Transformer v2.0). Простые базовые решения, такие как TF-IDF, случайное угадывание и мажоритарный классификатор, занимают последние три позиции в рейтинге. Несмотря на то что современные системы NLP соответствуют уровню человека или превосходят его в задачах разрешения лексической многозначности (RUSSE) и машинного чтения (MuSeRC, DaNetQA и RuCoS), все еще имеется точка роста на задачах разрешения кореференции (RWSD) и определения логической и причинно-следственной связи (RCB, TERRa, PARus).

3.2. MuSeRC and RuCoS: Наборы данных для задачи машинного чтения на материале русского языка.

Машинное чтение является одной из центральных задач NLU с широким прикладным применением, для решеня которой требуется общее понимание естественного языка, знания о мире и способность к логическому рассуждению. Были предложены различные формулировки задачи [137], включая заполнение пропуска в предложении, выбор одного из нескольких вариантов ответа, извлечение сегмента текста и ответ на вопрос в свободной форме. Однако, как и в

других исследовательских направлениях NLP, это направление сосредоточено на английском языке [87].

На момент выхода публикации Russian SuperGLUE задача машинного чтения на материале русского языка была преимущественно исследована в формулировке извлечения сегмента текста, в частности в рамках переноса знания между языками [34; 7; 23]. В связи с этим, мы предлагаем два новых набора данных на русском языке — MuSeRC и RuCoS — которые нацелены на оценку способности агрегировать информацию из нескольких предложений и использовать знания о мире для поиска ответа на вопрос. Дизайн MuSeRC и RuCoS аналогичен смежным наборам данных для английского языка MultiRC [51] и ReCoRD [121]. В данном разделе (i) более подробно описывается методология сбора и разметки текстовых данных для RuCoS, а так же методология оценки людей на этой задаче, (ii) представлен сравнительный статистический анализ ReCoRD и RuCoS, и (iii) представлены результаты эмпирической оценки людей и базовых решений на обеих задачах.

3.2.1. Метод

Сбор данных. Методология сбора текстовых данных для RuCoS включает в себя пять основных этапов: (1) сбор текстов из открытых новостных источников Lenta.ru[11] и Deutsche Welle[12], (2) генерация триплетов *<текст, предложение с пропуском, список вариантов ответа>*, (3) фильтрация примеров по частотности слов в текстах, (4) фильтрация примеров с помощью моделей машинного чтения, доступных в открытых библиотеках, и (5) фильтрация примеров с помощью разметчиков.

1. **Сбор текстов из новостных открытых источников**. Мы проводим автоматический сбор новостных статей из Lenta.ru и Deutsche Welle и извлекаем именованные сущности (англ. *named entities, NEs*) в статьях с помощью модели распознавания именованных сущностей на основе BERT, доступной в библиотеке DeepPavlov [18].
2. **Генерация триплетов**. Мы автоматически генерируем примеры набора данных в формате триплетов *<текст, предложение с пропуском, список вариантов ответа>*. Каждый *текст* состоит из первых нескольких абзацев новостной статьи и трех заголовков смежных новостных статей, которые отранжированы по значению косинусной близости между счетными векторными представлениями *текста* и заголовков. Заголовки предоставляют дополнительный контекст или резюмируют ситуацию, описанную в *тексте*. В качестве *предложения с пропуском* используется предложение, встречающееся после *текста*, содержащее как минимум одну упомянутую в *тексте* NE и удовлетворяющее критериям, описанным в работе [138]. Далее одна из NE в предложении заменяется на пропуск. В *список варантов ответа* входят все извлеченные из *текста* NEs.

[11] lenta.ru

[12] www.dw.com/ru/

3. **Фильтрация примеров по частотности**. Мы считаем частотность каждого текста как долю токенов, общее число употреблений которых превышает единицу на миллион слов Национального корпуса русского языка (НКРЯ)[13]. На данном этапе из набора данных удаляются примеры, частотность текста в которых составляет менее 70%.
4. **Фильтрация примеров с помощью моделей машинного чтения**. Мы проводим фильтрацию примеров с использованием моделей машинного чтения ruBERT и R-NET [121], доступных в библиотеке DeepPavlov. Пример удаляется из набора данных, если хотя бы одна из моделей правильно заполнила пропуск в предложении. После этой процедуры происходит формирование обучающей, валидационной и закрытой тестовой выборок, сбалансированных по новостному источнику.
5. **Фильтрация с помощью разметчиков**. Мы оцениваем качество полученных валидационной и закрытой тестовой выборок с помощью разметчиков на Толоке (см. инструкцию по разметке в § A.1 [37]). Проект по разметке данных включает в себя неоплачиваемое обучение, контрольные задания для контроля качества разметки и основную стадию разметки. Перед началом работы разметчику предоставляется подробная инструкция с описанием задачи и примерами разметки, которая доступна в любое время во время обучения и основной стадии разметки.

   Доступ к проекту предоставляется разметчикам, которые входят в топ-60% пользователей Толоки и успешно завершают обучение, правильно разметив как минимум 7 из 10 примеров. Задача разметчика состоит в том, чтобы (i) оценить связность между текстом и предложением с пропуском, (ii) сообщить, если ответ неочевиден или неоднозначен, (iii) выбрать все возможные варианты ответа и (iv) сообщить обо всех недочетах и ошибках, например, о неполной разметке NEs. Мы фильтруем размеченные примеры по времени ответа (> 30 секунд) и качеству выполнения контрольных заданий разметчиком (> 50%). После фильтрации голоса, полученные от трех до пяти разметчиков с помощью метода динамического перекрытия[14], агрегируются голосом большинства независимо для каждого варианта ответа. Двумя авторами данной публикации вручную проверяются примеры, в которых разметчики отметили недочеты или оставили какие-либо комментарии.

Общий статистический анализ. Для общего статистического анализа, результаты которого представлены в Таб. 5, используются токенизаторы spaCy для английского[15] и русского[16] языков. Распределение примеров по новостному источнику составляет 44%/56% в ReCoRD (CNN/Daily Mail News) и 67%/33% в RuCoS (Lenta.ru/Deutsche Welle). RuCoS включает в себя на

[13] ruscorpora.ru/

[14] toloka.ai/docs/dynamic-overlap

[15] github.com/explosion/spaCy

[16] github.com/aatimofeev/spacy_russian_tokenizer

Параметр	ReCoRD				RuCoS			
	Train	Dev	Test	Overall	Train	Dev	Test	Overall
Кол-во примеров	65,709	7,481	7,484	80,674	72,193	7,577	7,257	87,027
Кол-во предл. с пропуском	100,730	10,000	10,000	120,730	72,193	7,577	7,257	87,027
Кол-во уникальных предл. с пропуском	99,713	9,977	9,968	80,179	72,193	7,577	7,257	87,027
Кол-во уникальных текстов	65,258	7,133	7,279	79,670	72,193	7,577	7,257	87,027
|Словарь токенов в предл. с пропуском|	119,069	30,844	31,028	134,397	109,899	30,203	27,813	120,410
|Словарь токенов в текстах|	352,491	93,171	94,386	395,356	279,333	90,699	83,237	303,647
Кол-во токенов на предл. с пропуском	21.3	22.1	22.2	21.4	22.2	22.1	21.6	22.2
Кол-во токенов на текст	169.5	168.6	168.1	169.3	146.6	146.2	142.5	146.2
Кол-во NEs на текст	17.2	17.3	17.2	17.2	12.7	14.3	13.3	12.9
Частотность NE	7.1	4.4	4.3	7.5	8.9	5.0	5.3	9.6
Частотность правильного ответа	6.8	4.7	✗	6.5	10.2	4.1	✗	10.2
Доля высокочастотных токенов на предл. с пропуском	✗	✗	✗	✗	86.0	85.0	86.0	86.0
Доля высокочастотных токенов на текст	✗	✗	✗	✗	82.0	81.0	82.0	82.0

Таблица 5: Общий статистический анализ наборов данных ReCoRD и RuCoS.

6,3 тыс. больше примеров. В отличие от ReCoRD, в каждом примере RuCoS текст и предложение с пропуском уникальны, то есть каждому тексту сопоставляется только одно предложние с пропуском. Мы находим, что (i) тексты в RuCoS в среднем короче, (ii) предложения с пропуском в RuCoS содержат меньше NEs и (iii) ReCoRD более разнообразен относительно словарей токенов в текстах и предложениях с пропуском. Это можно объяснить особенностями языка, спецификой источников данных и распределением тем в новостных текстах. В то же время, RuCoS требует понимания богатой морфологии и высокой лексической вариативности в русском языке.

3.2.2. Эмпирическая оценка

Базовые решения. Мы проводим эмпирическую оценку базовых решений на основе счетных векторных представлений и языковой модели BERT: TF-IDF, ruBERT-base, mBERT и ruBERT-conv[17]. Подход к решению задач с помощью TF-IDF состоит из следующих этапов: (i) построение матрицы слово-документ на соответствующей обучающей выборке, (ii) замена пропуска в предложении каждым возможным вариантом ответа (RuCoS) или конкатенация текста с каждым возможным вариантом ответа, (iii) вычисление косинусной близости между счетными векторными представлениями текста и полученного предложения (RuCoS) или вопроса (MuSeRC), и (iv) предсказание в виде варианта ответа, который имеет максимальное значение близости. Кроме того, мы проводим оценку решения задач людьми на Толоке. Задача разметчиков заключается в том, чтобы (i) прочитать текст и предложение с пропуском, (ii) выбрать все возможные варианты ответа, которыми можно заполнить пропуск в предложении, и (iii) сооб-

[17] hf.co/DeepPavlov/rubert-base-cased-conversational

Модель	MuSeRC $F1_a$/EM	RuCoS F1/EM
TF-IDF	58.9/24.4	25.6/25.1
mBERT	66.8/33.6	30.6/29.6
ruBERT-conv	71.7/32.9	26.4/25.9
ruBERT-base	71.7/33.6	34.4/33.9
Человек	**80.6/42.0**	**93.0/92.4**

Таблица 6: Результаты эмпирической оценки базовых решений на наборах данных RuCoS и MuSeRC. **Метрики**: F1=F1-мера; $F1_a$=макро-усредненная F1-мера [51]; EM=полное совпадение.

щить о недочетах, если таковые имеются. Инструкции по разметке и примеры веб-интерфейса находятся в открытом доступе[18].

Метрики качества. Дизайн оценки аналогичен смежным работам [121; 51]. Полное совпадение (англ. *exact match*, EM) измеряет долю предсказаний, которые соответствуют всем правильным вариантам ответа (MuSeRC) или любому из правильных вариантов ответа (RuCoS). Макро-усредненная F1-мера (англ. *macro-average F1*, $F1_a$) является гармоническим средним точности и полноты, посчитанных по бинарной шкале для каждого варианта ответа. F1-мера (англ. *F1-score*, F1) измеряет пересечение между мешками слов предсказания модели и правильного ответа. F1-мера с максимальным значением среди всех возможных вариантов ответа нормируется на общее количество примеров в наборе данных.

Ключевые результаты. Таб. 6 представляет результаты проведенной эмпирической оценки. Моноязычные языковые модели достигают лучшего качества на задаче MuSeRC, в то время как mBERT превосходит по качеству ruBERT-conv на задаче RuCoS. ruBERT-base показывает лучшие результаты среди всех базовых решений на обеих задачах. Разница решения задачи людьми и базовыми решениями значительна, в особенности на задаче RuCoS.

3.2.3. Ретроспектива

Результаты эмпирической оценки указывают на то, что широко используемые языковые модели для русского языка на момент написания публикации значительно уступают людям в решении задач машинного чтения. В настоящее время системы NLP соответствуют уровню человека или превосходят его на MuSeRC и RuCoS (см. Таб. 4). К числу таких систем относятся дообученные языковые модели (FRED-T5, RuLeanALBERT) и ансамбли (Golden Transformer v2.0). Несмотря на это, предложенные наборы данных способствовали повышению репрезентативности русского языка, — который теперь занимает третье место по количеству ресурсов для задачи машинного чтения [87].

[18] github.com/RussianNLP/RussianSuperGLUE/HumanBenchmark

3.3. RuCoLA: Бенчмарк для оценки приемлемости предложения

Вопрос о приобретении языковыми моделями знания грамматики языка рассматривается с помощью оценки приемлемости предложения (англ. *acceptability judgments*), которая отражает, насколько предложение грамматично и естественно с точки зрения носителя языка [21]. В генеративной лингвистике данный подход широко используется для проверки гипотез в теориях формальной грамматики и усвоения языка [97]. Данные теоретические концепции адаптированы для оценки устойчивости языковых моделей [133], интерпретации качества на целевых задачах [17] и оценки грамматичности сгенерированных текстов [10; 11]. С момента выхода CoLA (Corpus of Linguistic Acceptability; [124]), бенчмарка для оценки приемлемости предложения на английском языке, сообщество NLP уделило внимание созданию смежных ресурсов на разных языках, за исключением русского [109; 41; 116; 130].

RuCoLA — это первый корпус предложений на русском языке, размеченных по бинарной шкале приемлемости. RuCoLA состоит из внутридоменных предложений, собранных вручную из публикаций и методических материалов по фундаментальной лингвистике, и внедоменных предложений, сгенерированных девятью моделями машинного перевода и генерации парафраза. Набор внедоменных предложений создан для прикладного применения оценки приемлемости предложения в задачах генерации русского языка. Мы предоставляем веб-сайт и открытый рейтинг систем для оценкии языковой компетенции предобученных моделей для русского языка.

3.3.1. Метод

Задача оценки приемлемости предложения формулируется как задача бинарной классификации, в которой от модели требуется определить, является ли входное предложение приемлемым.

- **Предложение**: "*Иван прилёг, чтобы он отдохнул.*"
- **Ответ**: `False` *(Неприемлемое предложение)*
- **Категория**: *Синтаксис*
- **Источник**: [107]

Сбор данных. RuCoLA включает в себя внутридоменную и внедоменную выборки (см. Таб. 7). Предложения и метки приемлемости для *внутридоменной* выборки собраны вручную из стандартных учебных пособий и академических публикаций по теоретической лингвистике, а так же методологических материалов по русской грамматике. Для *внедоменной* выборки предложения были сгенерированы с помощью девяти нейросетевых моделей, доступных в библиотеках EasyNMT[19] и russian-paraphrasers [36]. Для генерации использовались подвыборки наборов

[19] github.com/UKPLab/EasyNMT

Источник	Кол-во предл.	%	Тема
rusgram	563	49.7	Корпусная грамматика
[107]	1,335	73.9	Общий синтаксис
[63]	193	75.6	Синтаксические структуры
[70]	54	57.4	Генеративная грамматика
[76]	1,308	84.3	Семантика времени
[75]	1,374	90.8	Лексическая семантика
[77]	1,462	89.5	Отрицание
[100]	2,104	80.8	Семантика
[102]	1,444	36.6	Экзаменационные задания на грамматику
Внутридоменная выборка	**9,837**	**74.5**	
Машинный перевод	1,286	72.8	Перевод с англ. языка
Генерация парафраза	2,322	59.9	Автоматический парафраз
Внедоменная выборка	**3,608**	**64.6**	
Итог	**13,445**	**71.8**	

Таблица 7: Общая статистика бенчмарка RuCoLA по источнику. **Обозначения**: %=Доля приемлемых предложений; rusgram=коллекция методических материалов для проекта корпусного описания русской грамматики (доступ по ссылке: rusgram.ru).

данных, относящихся к разным предметным областям (Tatoeba [8], WikiMatrix [98], TED [83] и Yandex Parallel Corpus [5]), модели машинного перевода (OPUS-MT [108], M-BART50 [106] и M2M-100 [35]) и модели генерации парафраза разной емкости (ruGPT2-Large[20], ruT5[21] и mT5 [132]). Каждое сгенерированное предложение проходит двухэтапную процедуру разметки на Толоке.

Разметка данных. Каждый этап разметки сгенерированных предложений включает в себя неоплачиваемое обучение, контрольные задания для контроля качества разметки и основную стадию разметки. Перед началом работы разметчику предоставляется подробная инструкция с описанием задачи, описанием целевых классов и примерами разметки. Инструкции, примеры веб-интерфейса и другие детали проектов доступны в §B.1 и §B.2 [68].

Этап 1: Оценка приемлемости предложения. Первый этап разметки посвящен оценке приемлемости предложения. Доступ к проекту предоставляется разметчикам, которые входят в топ-60% пользователей Толоки и имеют сертификат о знании русского языка. Разметчик должен завершить обучение, правильно разметив как минимум 21 из 30 примеров. Голоса от разметчиков, которые успешно справились менее чем с половиной контрольных заданий, исключаются. Голоса, полученные от трех до пяти оставшихся разметчиков с помощью метода динамического перекрытия, агрегируются голосом большинства.

[20] hf.co/ai-forever/ruGPT2-large

[21] hf.co/cointegrated/rut5-base-paraphraser

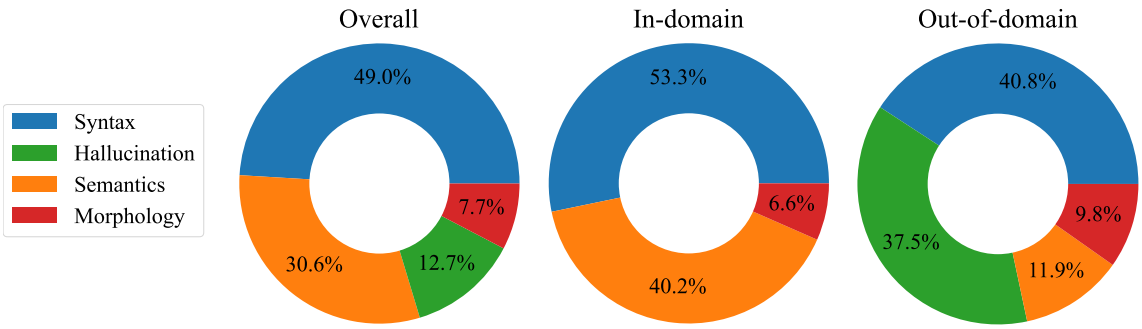

Рис. 1: Распределение категорий ошибок в наборе неприемлемых предложениях.

Этап 2: Разметка по категориям ошибок. Второй этап нацелен на валидацию и разметку предложений, которым на предыдущем этапе присвоился класс "неприемлемое предложение", по следующим верхнеуровневым категориям ошибок: "Морфология", "Синтаксис", "Семантика", "Галлюцинации" и "Другое". В этом проекте мы собираем команду разметчиков из 30 студентов бакалаврских и магистерских программ по филологии и лнгвистике из нескольких российских университетов. Мы проводим онлайн-семинар для обсуждения проекта разметки и смежных работ [124; 42; 140]. Каждый студент проходит обучение на 15 примерах и имеет возможность общаться с авторами публикации в групповом чате. После фильтрации размеченных примеров по времени ответа (> 30 секунд на 5 предложений) голоса разметчиков агрегируются голосом большинства независимо для каждого варианта ответа. Предложения, которым присвоилось несколько классов категорий ошибок или класс "Другое", исключаются.

Лингвистические феномены. Каждое неприемлемое предложение относится к одной из верхнеуровневых категорий ошибок: морфология, синтаксис, семантика и галлюцинации. Лингвистические феномены, относящиеся к этим категориям, хорошо представлены в теоретической и корпусной лингвистике, а так же специфичны для генеративных моделей. Примеры таких явлений включают неправильное словообразование, нарушения согласования, нарушение порядка слов, неправильное использование отрицания, бессмысленные предложения, избыточные повторы частей предложения, артефакты стратегий декодирования и другие (см. § A.1 и § A.2 [68]).

Общий статистический анализ. Набор предложений фильтруется по диапазону от 4 до 30 токенов включительно с помощью токенизатора razdel[22]. Среднее количество токенов в предлоежениии составляет 11. Как и в §3.2.1, мы оцениваем частотность каждого предложения с использованием НКРЯ. Доля высокочастотных токенов в предложении составляет в среднем 92%. Рис. 1 иллюстрирует распределение категорий ошибок в неприемлемых предложениях. Синтаксические нарушения составляют 53,3% и 40,8% во внутридоменной и внедоменной выборках. Внутридоменные предложения содержат 40,2% семантических и 6,6% морфологических нару-

[22] github.com/natasha/razdel

шений, в то время как внедоменные — 11,9% и 9,8% соответственно. 12,7% нарушений от общего количества неприемлемых предложений относятся к галлюцинациям.

Набор внутридоменных предложений разделен на обучающую, валидационную и закрытую тестовую выборки в соотношении 80/10/10 (7,9 тыс./1 тыс./ 1 тыс. примеров). Набор внедоменных предложениий разделен на валидационную и закрытую тестовую выборки в соотношении 50/50 (1,8 тыс./1,8 тыс. примеров).

3.3.2. Эмпирическая оценка

Базовые решения. Мы проводим эмпирическую оценку трех типов базовых решений: наивные модели (TF-IDF и мажоритарный классификатор), вероятностные меры приемлемости предложения (PenLP) и широкий набор моноязычных и мультиязычных языковых моделей (ruBERT-base[23], ruRoBERTa-large[24], ruT5-base[25], XLM-R-base [29] и RemBERT [22]). Под TF-IDF подразумевается логистическая регрессия, обученная на N-граммных признаках TF-IDF. Значение N $\in [1; 3]$; размерность вектора признаков равна 2,5 тыс. Мера PenLP [57] посчитана с использованием ruGPT3-medium[26] как сумма лог-вероятностей токенов в предложении $P(s)$, отнормированная на длину предложения в токенах с коэффициентом α (см. уравнение Equation 1). Целевой класс определяется с помощью порога значения меры, посчитанного методом кросс-валидации на обучающей и валидационной выборках.

$$\text{PenLP}(s) = \frac{P(s)}{((5 + |s|)(5 + 1))^{\alpha}} \tag{1}$$

Мы так же проводим оценку решения задачи людьми на всей внутридоменной закрытой тестовой выборке и половине внедоменной закрытой тестовой выборки. Дизайн проекта аналогичен первому этапу разметки предложений, описанному в §3.3.1, с несколькими отличиями: (i) вариант ответа "Не знаю" удален, (ii) разметчиками являются 16 студентов бакалаврских и магистерских программ по филологии и лнгвистике, и (iii) голоса разметчиков агрегируются с помощью метода Дэвида-Скина [31], доступном напрямую через веб-интерфейс Толоки. Среднее качество выполнения контрольных заданий составляет 75%.

Метрики качества. Качество оценки приемлемости предложения по бинарной шкале определяется с помощью accuracy (Acc.) и коэффициента корреляции Мэтьюса (англ. *Matthews Correlation Coefficient*, MCC [65]). MCC является целевой метрикой. Базовые решения обучены или дообучены с помощью подбора гиперпараметров на валидационной выборке. Описание

[23] hf.co/ai-forever/ruBERT-base

[24] hf.co/ai-forever/ruRoBERTa-large

[25] hf.co/ai-forever/ruT5-base

[26] hf.co/ai-forever/ruGPT3-medium

Базовое решение	Общий рез-т		Внутридоменный набор		Внедоменный набор	
	Acc.	MCC	Acc.	MCC	Acc.	MCC
Наивные модели						
Мажоритарный кл.	68.05 ± 0.0	0.0 ± 0.0	74.42 ± 0.0	0.0 ± 0.0	64.58 ± 0.0	0.0 ± 0.0
Linear	67.34 ± 0.0	0.04 ± 0.0	75.53 ± 0.0	0.17 ± 0.0	62.86 ± 0.0	-0.02 ± 0.0
Вероятностные меры приемлемости предложения						
ruGPT-3	55.79 ± 0.0	0.27 ± 0.0	59.39 ± 0.0	0.19 ± 0.0	53.82 ± 0.0	0.30 ± 0.0
Моноязычные языковые модели						
ruBERT	75.9 ± 0.42	0.42 ± 0.01	78.82 ± 0.57	0.4 ± 0.01	74.3 ± 0.71	0.42 ± 0.01
ruRoBERTa	<u>80.8</u> ± 0.47	<u>0.54</u> ± 0.01	<u>83.48</u> ± 0.45	<u>0.53</u> ± 0.01	<u>79.34</u> ± 0.57	<u>0.53</u> ± 0.01
ruT5	71.26 ± 1.31	0.27 ± 0.03	76.49 ± 1.54	0.33 ± 0.03	68.41 ± 1.55	0.25 ± 0.04
Мультиязычные языковые модели						
XLM-R	65.73 ± 2.33	0.17 ± 0.04	74.17 ± 1.75	0.22 ± 0.03	61.13 ± 2.9	0.13 ± 0.05
RemBERT	76.21 ± 0.33	0.44 ± 0.01	78.32 ± 0.75	0.4 ± 0.02	75.06 ± 0.55	0.44 ± 0.01
Человек	**84.08**	**0.63**	**83.55**	**0.57**	**84.59**	**0.67**

Таблица 8: Результаты оценки приемлемости предложения по бинарной шкале. **Метрики**: Acc.=доля правильных ответов; EM=полное совпадение; MCC=Коэффициент корреляции Мэтьюса [65]. Значения, выделенные жирным начертанием, означают лучший результат, и подчеркнутые значения — второй лучший результат.

результатов эмпирической оценки ниже проводится по значениям метрик качества на закрытой тестовой выборке, усредненным по 10 экспериментальным запускам с разными значениями, определяющими состояние генератора случайных чисел (англ. *random seed*).

Оценка приемлемости предложения по бинарной шкале

Ключевые результаты. В Таб.8 представлены результаты эмпирической оценки. Основные выводы заключаются в следующем: (i) ruRoBERTa и RemBERT достигают лучших результатов среди базовых решений, (ii) качество наивных моделей близко к нулю, (iii) лучшие языковые модели хорошо обобщаются на внедоменные предложения с разницей до 0,04 MCC, и (iv) качество оценки приемлемости внедоменных предложений людьми выше, чем внутридоменных, что свидетельствует о том, что неприемлемость сгенерированных предложений легче определить.

Анализ ошибок. Результаты количественного анализа 250 неправильно классифицированных предложений показывают, что (i) языковые модели часто оценивают неграмматичные предложения с морфологическими и синтаксическими нарушениями как приемлемые, (ii) люди допускают ошибки в длинных предложениях со сравнительными и подчинительными клаузами, а так же предложным управлением, и (iii) большинство языковых моделей достигает

высокое значение полноты (англ. *recall*) на предложениях с галлюцинациями, что подтверждает потенциал прикладного применения RuCoLA.

Влияние длины предложения. Ключевым выводом анализа влияния длины предложения на целевое качество является то, что качество моделей консистентно относительно всех групп длин. Но несмотря на то что качество моделей на предложениях разных длин сильно варьируется и ухудшается по мере увеличения длины, разметчики превосходят языковые модели на всех группах длин. Данные результаты согласуются с работой [122].

Перенос знания между языками

RuCoLA способствует проведению исследований по качеству переноса знания грамматики между разными языками. Мы проводим следующий эксперимент: мультиязычные модели дообучаются на обучающей выборке одного языка и оцениваются на тестовой выборке другого языка. В эксперименте используются смежные бенчмарки на материале английского (CoLA) и итальянского языков (ItaCoLA; [109]) и четыре мультиязычные языковые модели: mBERT, XLM-R-base, XML-R-large и RemBERT.

Ключевые результаты. Мы находим, что (i) разница в целевом качестве в зависимости от исходного языка незначительна, (ii) языковые модели, дообученные в моноязычной конфигурации, значительно превосходят языковые модели, дообученные в конфигурации переноса знания между языками, что согласуется с результатами смежной работы [109], (iii) RemBERT показывает наилучшие результаты в обеих конфигурациях среди мультиязычных языковых моделей, и (iv) RemBERT и XLM-R-large хорошо обобщаются на внедоменный набор предложений RuCoLA, в частности при дообучении на английском и итальянском языках.

3.3.3. Ретроспектива

Открытый рейтинг систем RuCoLA[27] показывает, что на текущий момент предложенная задача является сложной для существующих языковых моделей. На открытом рейтинге систем представлено более 30 различных подходов к оценке приемлемости предложения, включая дистилированные языковые модели[28] и классификаторы на основе признаков TDA [82]. Бенчмарк RuCoLA поспособствовал практическому использованию оценки приемлемости предложения, например, для фильтрации неприемлемых автоматических переводов подписей к изображениям при сборе предобучающего корпуса ruDALL-E[29].

[27] rucola-benchmark.com/leaderboard

[28] hf.co/cointegrated/ruBERT-tiny

[29] hf.co/ai-forever/ruDALL-E-Malevich

3.4. RuATD: Соревнование по автоматическому распознаванию сгенерированных текстов

Современные языковые модели способны генерировать высококачественные тексты на многих языках. Однако такие модели могут использоваться в злоумышленных целях [126], например, для генерации заведомо ложных новостей [136] и экстремистских материалов [66]. Область распознавания сгенерированного текста (англ. *artificial text detection*, ATD) направлена на разработку ресурсов и вычислительных методов для снижения рисков злонамеренного использования генеративных языковых моделей.

Чтобы исследовать данную проблему на материале русского языка, мы представляем бенчмарк RuATD, предложенный в рамках кампании Dialogue Evaluation[30] в 2022 году как открытое соревнование. Участникам соревнования предлагается решить две задачи, сформулированные по аналогии с тестом Тюринга [111] и задачи определения автора текста [112]. RuATD покрывает широкий спектр предметных областей, генеративных языковых моделей и задач генерации естественного языка. Мы предоставляем два открытых рейтинга систем на платформе Kaggle, которые остаются общедоступными после завершения соревнования.

3.4.1. Метод

Распознавание сгенерированного текста — это задача бинарной классификации, направленная на прогнозирование, был ли входной текст сгенерирован нейросетевой моделью или написан человеком.

- **Текст**: *"Я был готов помочь ему в опасности своей жизни."*
- **Ответ**: Модель

Определение автора текста — это задача, направленная на прогнозирование автора входного текста. Задача формулируется как задача многоклассовой классификации с 14 целевыми классами: человек и 13 генеративных моделей.

- **Текст**: *"Я был полон решимости помочь ему, даже рискуя собственной жизнью."*
- **Ответ**: Человек

Сбор данных. RuATD состоит из 215 тыс. текстов, написанных человеком, и текстов, сгенерированных с помощью 13 нейросетевых моделей для русского языка. Методология создания RuATD включает три основных этапа: (i) сбор текстов, написанных человеком, (ii) генерация текстов и (iii) фильтрация текстовых данных. Мы приводим краткое описание методологии ниже и отсылаем читателя к разделу §2 в [101] для получения подробной информации о спецификации генеративных моделей и данных, гиперпараметрах дообучения моделей и гиперпараметрах генерации текста.

[30] www.dialog-21.ru/en/evaluation

1. **Сбор текстов, написанных человеком**. Мы проводим автоматический сбор естественных текстов из открытых источников, относящимся к нескольким предметным областям: НКРЯ, посты в социальных сетях, статьи из Википедии (топ-100 самых просматриваемых страниц в 2016-2021 гг.), новостные статьи (Lenta.ru, Комсомольская правда, Интерфакс, Известия и др.), оцифрованные дневники [67] и корпус Минэкономразвития РФ, содержащий программы стратегического развития [47]. Мы также включаем золотой стандарт, собранный из наборов данных для машинного перевода WikiMatrix и Tatoeba.
2. **Генерация текстов**. Мы используем тексты, написанные человеком, в качестве входных данных для генеративных моделей, дообученных на одной или нескольких задач генерации естественного языка: машинного перевода, генерации парафраза, упрощения текста и суммаризации текста. Кроме того, мы используем методы обратного перевода и открытой генерации.

    - **Машинный и обратный перевод**. Мы автоматически переводим подвыборки наборов данных Tatoeba и WikiMatrix на трех языковых парах (английский-русский, французский-русский, испанский-русский) тремя моделями, используя библиотеку EasyNMT: OPUS-MT, M-BART50 и M2M-100. При генерации текста методом обратного перевода предложение на русском языке переводится на целевой язык, а затем переводится обратно на русский язык.
    - **Генерация парафраза**. Мы генерируем паррафразы с помощью моделей, доступных в библиотеке russian-paraphrasers: ruGPT2-large, ruT5-base и mT5 версии Small и Large.
    - **Упрощение текста**. Для генерации текстов в рамках задачи упрощения текста мы дообучаем ruGPT3-small, ruGPT3-medium, ruGPT3-large, mT5-large и ruT5-large на наборе данных RuSimpleSentEval [93].
    - **Суммаризация текста**. Мы используем ruT5-base и M-BART, дообученные на наборе данных Gazeta [40].
    - **Открытая генерация**. В методе открытой генерации моделям ruGPT3-small, ruGPT3-medium и ruGPT3-large на вход подается начало естественного текста, для которого генерируется продолжение.

3. **Фильтрация текстовых данных**. Мы используем набор эвристик и общих статистических параметров, чтобы (i) удалить текстовые дубликаты, скопированные входные тексты и пустые выходы нейросетевых моделей, (ii) исключить тексты, содержащие нецензурную лексику, (iii) отфильтровать переводы с помощью инструмента автоматического распознавания языка[31] и (iv) отфильтровать оставшиеся тексты по диапазонам длин в токенах: 5-25 токенов (машинный и обратный перевод, генерация парафраза), 10-30 токенов (упрощение текста), 15-60 токенов (суммаризация текста) и 85-400 токенов (открытая генерация).

[31] github.com/fedelopez77/langdetect

Задача	Модель	Кол-во текстов	N	%	Предметная область	Задача	Модель	Кол-во текстов	N	%	Предметная область
Обратный перевод	Человек M-BART50 M2M-100 OPUS-MT	35,588	12.9	88.0	НКРЯ, Википедия, новостные статьи, дневники, WikiMatrix, Tatoeba, Минэк	**Машинный перевод**	Человек M-BART50 M2M-100 OPUS-MT	35,860	11.5	89.0	WikiMatrix, Tatoeba
Открытая генерация	Человек ruGPT3-small ruGPT3-medium ruGPT3-large	37,499	141.5	85.0	НКРЯ, Википедия, новостные статьи, дневники, Минэк, соц. сети	**Суммаризация текста**	Человек M-BART M-BART50 ruT5-base	17,164	33.5	86.0	НКРЯ, Википедия, новостные статьи, дневники, Минэк
Генерация парафраза	Человек mT5-small mT5-large ruGPT2-large ruGPT3-large ruT5-base	44,298	13.0	85.0	НКРЯ, Минэк, соц. сети, Википедия, новостные статьи, дневники	**Упрощение текста**	Человек mT5-large ruGPT3-small ruGPT3-medium ruGPT3-large ruT5-large	44,700	18.3	86.0	НКРЯ, Минэк, соц. сети, Википедия, новостные статьи, дневники

Таблица 9: Общий статистический анализ бенчмарка RuATD. **Обозначения**: N=среднее количество токенов в тексте; %=доля высокочастотных токенов в тексте; Минэк=корпус Минэкономразвития РФ.

Общий статистический анализ. В Таб. 9 представлены результаты общего статистического анализа бенчмарка RuATD по задаче генерации естественного языка, генеративной модели и предметной области. Среднее количество токенов в тексте составляет 37,9, в зависимости от задачи генерации. Средняя доля высокочастотных токенов примерно одинакова в естественных и сгенерированных текстах: 86% и 87% соответственно. Собранный корпус разделен на четыре выборки в соотношении 60/10/15/15: обучающая (130 тыс. примеров), валидационная (21 тыс. примеров), открытая тестовая (32 тыс. примеров) и закрытая тестовая (32 тыс. примеров). Открытая тестовая выборка доступна во время всего соревнования и позволяет участникам разрабатывать и улучшать свои решения. Закрытая тестовая выборка определяет окончательное место участников в рейтинге систем и предотвращает переобучение на открытой тестовой выборке. Обучающая, валидационная, открытая и закрытая тестовая выборки используются в обеих формулировках задач с единственным отличием в целевых классах. Целевой класс "Модель", используемый в задаче бинарной классификации, декомпозируется на 13 названий генеративных моделей в задаче многоклассовой классификации.

3.4.2. Эмпирическая оценка.

Базовые решения. Участникам соревнования предоставляется два базовых решения: TF-IDF и ruBERT-base. Под TF-IDF подразумевается логистическая регрессия, обученная на N-граммных признаках TF-IDF. Значение $N \in [1;3]$; размерность вектора признаков снижена с помощью сингулярного разложения матрицы до 5 тыс. ruBERT-base дообучается на соответствующей задаче. Мы также проводим оценку решения задачи распознавания сгенерированного текста людьми на Толоке, используя стратифицированные подвыборки из 2,5 тыс. примеров открытой тестовой и закрытой тестовой выборок. § A в [101] содержит инструкцию по разметке. Мы предоставляем доступ к проекту разметчикам, входящим в топ-70% пользователей Толоки. Каждый разметчик должен пройти обучение, правильно ответив на не менее чем 27 из 32 примеров. Мы фильтруем размеченные примеры по времени ответа (> 15 секунд на пять текстов) и качеству выполнения контрольных заданий разметчиком (> 50%). Оставшиеся

Ранк	Распознавание сгенерированного текста		Определение автора текста	
	Команда	Acc.	Команда	Acc.
1	MSU	0.829	Posokhov Pavel	0.650
2	Igor	0.827	Yixuan Weng	0.647
3	Orzhan	0.826	Orzhan	0.646
4	mariananieva	0.824	MSU	0.628
5	Ivan Zakharov	0.822	ruBERT baseline	0.598
6	Yixuan Weng	0.818	Nikita Selin	0.590
7	ilya koziev	0.817	Victor Krasilnikov	0.550
8	miso soup	0.811	Petr Grigoriev	0.458
9	Eduard Belov	0.810	TF-IDF baseline	0.443

Таблица 10: Топ-девять позиций на двух открытых рейтингах систем RuATD. **Метрика**: Acc.=доля правильных ответов.

голоса, полученные от трех до пяти разметчиков с помощью метода динамического перекрытия, агрегируются голосом большинства.

Метрики качества. В качестве целевой метрики используется accuracy.

Ключевые результаты. Мы излагаем основные результаты оценки базовых решений и 38 решений от участников соревнования (см. Таб. 10): (i) качество систем зависит от длины текста (чем длиннее текст, тем выше доля правильных ответов), (ii) задача определения автора текста не тривиальна, что потенциально означает, что тексты, написанные человеком, и тексты, сгенерированные различными моделями, имеют похожие признаки, и (iii) люди уступают системам в распознавании сгенерированного текста на 0,169 accuracy. Последний тезис согласуется с результатами смежных работ на материале английского языка [50; 113].

3.4.3. Ретроспектива

Область ATD остается сосредоточенной на трех языках: английском, китайском и русском. Бенчмарки ATD становятся сложнее и покрывают большое количество генеративных языковых моделей, предметных областей, стратегий декодирования и методов генерации естественного текста. Быстрое распространение генеративных языковых моделей требует непрерывного обновления бенчмарков ATD и разработки более устойчивых детекторов сгенерированного текста, поскольку их качество распознавания уменьшается по мере увеличения количества параметров модели-генератора [104]. В разделе §3.5, мы представляем новый метод для распознавания сгенерированного текста, который превосходит базовые решения и лучше обобщается при распознавании текстов, сгенерированных моделями GPT-2, которые отсутствовали в обучающей выборке.

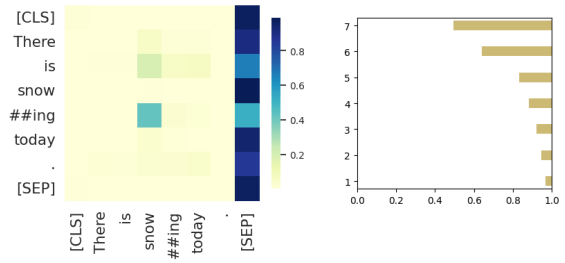

Рис. 2: Пример карты внимания (слева) и баркода (справа) [20].

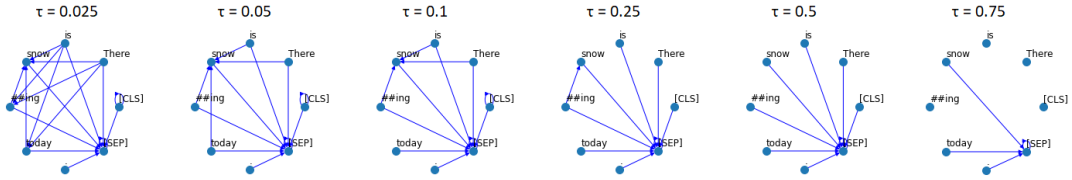

Рис. 3: Пример процедуры фильтрации [20].

3.5. Распознавание сгенерированного текста с помощью топологического анализа карт внимания

С последними достижениями в области генерации естественного языка проблема низкого качества распознавания сгенерированного текста человеком становится более актуальной. В области ATD были предложены различные вычислительные методы для человеко-машинного взаимодействия при решении данной задачи, такие как линейные классификаторы на счетных векторных представлениях и лингвистических признаках [46; 9], извлечение статистических признаков текста, посчитанных с помощью языковой модели [39; 33] и дообучение языковой модели [104]. Последний подход наиболее эффективен, но не обладает интерпретируемостью и устойчивостью к отсутствовавшим в обучающей выборке генеративным моделям [48].

Мы представляем гибридный детектор сгенерированного текста, который объединяет преимущества детекторов на основе признаков и языковых моделей: интерпретируемость и высокие показатели качества. Мы предпринимаем одну из первых попыток адаптировать методы прикладной топологии к механизму внимания языковых моделей. Наш подход включает в себя (i) извлечение трех типов признаков TDA из графового представления карты внимания языковой модели и (ii) обучение линейного классификатора, используя конкатенацию полученных признаков.

3.5.1. Метод

Карта внимания языковой модели (см. Рис. 2; слева) представлена как взвешенный граф, вершины которого соответствуют токенам в тексте. Веса механизма внимания поставлены в соответствие весам ребер графа. Данное графовое представление используется для процедуры *"фильтрации"* (см. Рис. 3), — построения множества графов внимания, упорядоченного по

Предметная область	Генеративная модель	|Train|		|Dev|		|Test|		|Словарь токенов|		Длина	
		Ч	М	Ч	М	Ч	М	Ч	М	Ч	М
Reddit	GPT-2-small жадный поиск	20 тыс.	20 тыс.	2.5 тыс.	2.5 тыс.	2.5 тыс.	2.5 тыс.	220 тыс.	532 тыс.	593 ± 177	515 ± 322
Отзывы на продукты Amazon	GPT-2-XL жадный поиск	5 тыс.	5 тыс.	1 тыс.	1 тыс.	4 тыс.	4 тыс.	47 тыс.	49 тыс.	179 ± 170	177 ± 171
Новостные статьи	GROVER ядерная генерация	5 тыс.	5 тыс.	1 тыс.	1 тыс.	4 тыс.	4 тыс.	98 тыс.	75 тыс.	721 ± 636	519 ± 203

Таблица 11: Общий статистический анализ наборов данных, используемых в экспериментах по распознаванию сгенерированного текста. **Обозначения**: **Ч**=человек; **М**=модель.

возрастающему порогу веса внимания τ_i. Фильтрация ребер, вес которых меньше заданного порога, влияет на структуру графа и его свойства. Методы TDA позволяют отслеживать такие изменения, определяя моменты появления (*"рождения"*) или исчезновения (*"смерти"*) того или иного свойства. Время жизни свойства представляется как набор интервалов, называемых *"баркодом"* (см. Рис.2; справа), в котором каждый интервал (*"бар"*) длится от рождения свойства до его смерти. Баркод характеризует стабильность свойств графа.

Мы извлекаем три типа признаков TDA из карт внимания языковой модели. Признаки вычисляются по набору заранее определенных пороговых значений для каждой головы внимания и далее конкатенируются.

1. **Топологические признаки**. Топологические признаки включают в себя нулевое число Бетти β_0 и первое число Бетти β_1 неориентированного графа и стандартные свойства ориентированного графа: количество сильно связанных компонент, ребер и циклов.
2. **Описательные характеристики баркодов**. Мы вычисляем $0/1$-мерные баркоды из графа внимания и их описательные характеристики: сумму/среднее/дисперсию длин баров, количество баров с временем рождения/смерти больше/меньше заданного порога, и энтропию баркодов.
3. **Признаки на основе расстояния до паттернов внимания**. Форма графов внимания следует различным паттернам внимания: внимание к предыдущему/текущему/следующему токену, внимание к специальным токенам [SEP]/[CLS] и внимание к пунктуационным знакам [24]. На Рис. 2 (слева) показан пример паттерна внимания к специальному токену [SEP]. Мы представляем паттерны внимания в виде бинарных матриц и вычисляем Евклидову норму разности данных матриц, отнормированную на сумму их Евклидовых норм.

3.5.2. Эмпирическая оценка

Наборы данных. Мы проводим эксперименты на трех наборах данных из трех предметных областей (см. Таб. 11): (i) Reddit & GPT-2-small [84], (ii) отзывы на продукты Amazon & GPT-2-XL [4; 104] и (iii) новостные статьи & GROVER [136].

Модель	Reddit & GPT-2 Small	Отзывы на продукты Amazon & GPT-2 XL	Новостные статьи & GROVER
TF-IDF, N-grams	68.1	54.2	56.9
BERT [CLS trained]	77.4	54.4	53.8
BERT [Fully finetuned]	**88.7**	**60.1**	**62.9**
BERT [SLOR]	78.8	59.3	53.0
Топологические признаки	86.9	59.6	63.0
Описательные характеристики баркодов	84.2	60.3	61.5
Признаки на основе р. до паттернов внимания	85.4	61.0	62.3
Конкатенация признаков TDA	**87.7**	**61.1**	**63.6**

Таблица 12: Результаты эмпирической оценки детекторов сгенерированного текста. Значения целевой метрики умножены на 100.

Базовые решения. Базовые решения на основе языковой модели BERT[32] включают в себя: (i) BERT [CLS trained] — это линейный слой, обученный на векторном представлении специального токена [CLS] (веса BERT заморожены), (ii) BERT [Fully finetuned] дообучен на целевой задаче. Кроме этого, мы обучаем логистическую регрессию на следующих признаках: (iii) N-граммные признаки TF-IDF; значение $N \in [1, 2]$, и (iv) BERT [SLOR] [80], значение вероятностной меры приемлемости, посчитанной как псевдоперплексия [57].

Предложенный детектор. Мы обучаем несколько конфигураций логистической регрессии на признаках TDA, извлеченных из матриц внимания вышеупомянутой языковой модели BERT: (i) *топологические признаки*, (ii) *описательные характеристики баркодов*, (iii) *признаки на основе расстояния до паттернов внимания*, и (iv) *конкатенация признаков TDA*.

Метрики качества. В качестве целевой метрики используется accuracy.

Распознавание сгенерированного текста

Ключевые результаты. В Таб. 12 представлены результаты проведенной эмпирической оценки. Мы находим, что предложенный метод (i) превосходит базовые решения на основе счетных векторных представлений и языковой модели BERT на 10% accuracy и (ii) достигают качества, сравнимого с дообученной языковой моделью BERT.

Анализ устойчивости к генеративным моделям, отсутствующим в обучающей выборке. В данной экспериментальной постановке детекторы обучены на естественных текстах и текстах, сгенерированных моделью GPT-2-small, и используются для распознавания текстов, сгенерированных моделями GPT-2 большей емкости: GPT-2-medium, GPT-2-large и GPT-2-xl. Ключевой результат заключается в том, что детектор, обученный на *топологических признаках*

[32] hf.co/bert-base-uncased

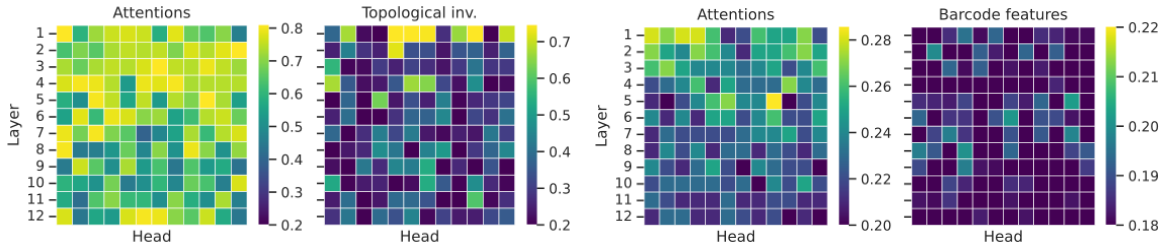

Рис. 4: Результаты диагностического тестирования голов внимания на задаче прогнозирования длины предложения (слева) и глубины синтаксического дерва (справа). **Обозначения:** Attentions=веса внимания языковой модели BERT. OX=порядковый номер головы внимания. OY=порядковый номер слоя. Topological inv.=топологические признаки. Чем ярче цвет, тем выше значение accuracy.

более устойчив, чем базовые решения, однако незначительно уступает дообученной языковой модели BERT на тестовой выборке, включающей тексты GPT-2-small.

Диагностическое тестирование

Ключевые результаты. На Рис. 4 представлены результаты диагностического тестирования голов внимания модели BERT и топологических признаков на двух задачах: прогнозирование длины предложения и глубины синтаксического дерева. Основные результаты заключаются в следующем: топологические признаки (i) чувствительны к поверхностным и синтаксическим признакам (ii) могут терять информацию о свойствах, что характеризуется разным качеством одних и тех же голов внимания, и (iii) не кодируют семантические свойства (например, временные формы глагола). Однако этой информации достаточно для рассматриваемой задачи распознавания сгенерированного текста.

3.5.3. Ретроспектива

Методы TDA нашли широкое применение в различных задачах МО, таких как распознавание изображений [105], сегментация изображений [25], классификация текстов [96; 127] и оценка генерации естественного языка [27]. Результаты нескольких последующих работ демонстрируют, что предложенная методология может быть адаптирована для достижения новых передовых результатов на задачах обработки речи [110] и достижения уровня человека на задачах оценки приемлемости предложения [20].

3.6. Vote'n'Rank: Пересмотр методологии эталонного тестирования с помощью теории многокритериального принятия решений

Использование среднего арифметического в качестве метода агрегации результатов эталонного тестирования подвергается сомнению, поскольку данный метод: (i) предполагает, что целевые метрики однородны [28], (ii) не учитывает сложность задач [69], (iii) полагается на абсо-

лютную разницу в целевых метриках [81] и (iv) способствует смещению оценки, в связи с тем что системы могут превосходить друг друга лишь на определенных, а не на всех задачах [1].

Мы представляем Vote'n'Rank — фреймворк для ранжирования систем NLP на бенчмарках, включающих в себя множество задач, или в сценариях эталонного тестирования, учитывающих несколько критериев оценки. Фреймворк позволяет решить проблемы, связанные с применением среднего арифметического для агрегации результатов, с помощью принципов теории многокритериального принятия решений [2]. В данном разделе представлено описание предложенных методов агрегации результатов, экспериментального дизайна и ключевых результатов.

3.6.1. Метод

Теория многокритериального принятия решений оперирует следующими понятиями: (i) избиратель или критерий (англ. *voter*; в случае анализа бенчмарков — метрика качества) и (ii) альтернатива (англ. *alternative*; в случае анализа бенчмарков — оцениваемая система МО). Мы преследуем две основные цели процедуры агрегации: выбор победителя или лучшей альтернативы и ранжирования альтернатив в порядке убывания по заданным критериям.

Методы агрегации результатов. Методы агрегации результатов Vote'n'Rank относятся к трем классам правил голосования (англ. *voting rules*): скоринговые правила (*правило большинства, правило Борда и правило Даудолла*), итеративные скоринговые правила (*пороговое правило* и *правило Болдуина*) и мажоритарные правила (*правило Кондорсе*, *правило Коупленда*, и *правило минимакса*). Дизайн правил основан на теоретических основах теории многокритериального принятия решений и общепринят в мировом сообществе [3; 6; 74]. Ниже мы предоставляем рекомендации по выбору метода агрегации и отсылаем читателя к§2 и § A в [85] для примеров и более подробной информации о свойствах методов агрегации.

- *Правило большинства* стоит использовать, если нужно учитывать только лучшие альтернативы по каждому критерию.
- *Правило Борда* и *правило Даудолла* наиболее подходящие для случаев, когда важны все позиции в рейтинге систем. При этом *правило Даудолла* присваивает более высокий вес верхним позициям в рейтинге систем.
- *Пороговое правило* полезно в тех случаях, когда пользователь хочет минимизировать вклад малозначимых критериев.
- Пользователю рекомендуется использовать правила *Болдуина, Кондорсе, Коупленда* или *Минимакса*, если требуется определить *победителя Кондорсе*, — альтернативы, которая доминирует над всеми другими альтернативами при попарном сравнении. Основное отличие заключается в поведении правил в случае отсутствия такой альтернативы. В частности, *правило Болдуина* выбирает альтернативу, которая остается после исключения всех других в соответствии с *правилом Борда. Правило Коупленда* выбирает альтернативу, которая доминирует

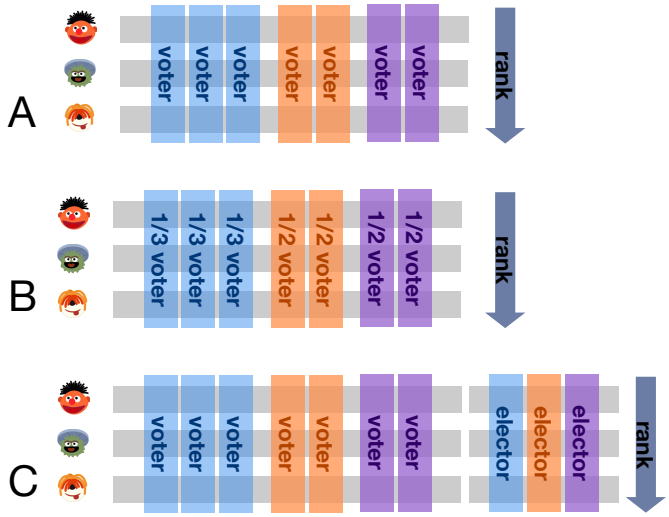

Рис. 5: Три способа использования Vote'n'Rank. A: базовая агрегация. B: взвешенная агрегация. C: двухэтапная агрегация.

над другими в большинстве случаев и меньше всех проигрывает другим альтернативам. *Правило Минимакса* выбирает альтернативу с минимальным количеством поражений при попарном сравнении.

Фреймворк. Рис. 5 иллюстрирует три способа использования Vote'n'Rank для агрегации результатов эталонного тестирования. Допустим, бенчмарк имеет три оцениваемых альтернативы и состоит из семи избирателей, сгруппированных по задачам, например: машинное чтение, классификация текстов и вопросно-ответные задачи.

A. Базовая агрегация: метод агрегации применяется к бенчмарку в исходном виде.

B. Взвешенная агрегация: каждому избирателю в группе присваивается групповой вес, равный $1/|T_{group}|$. Вес синей группы равен $1/3$, а вес оранжевой и фиолетовой группы равен $1/2$. Каждая группа вносит одинаковый вклад в итоговый рейтинг, независимо от количества избирателей.

C. Двухэтапное агрегирование: каждая группа избирателей рассматривается как отдельный рейтинг систем. Мы независимо применяем метод агрегации к каждой группе избирателей и определяем промежуточный рейтинг, обозначенный в примере как «избиратель » . Затем мы агрегируем рейтинги систем внутри каждой группы, повторно применяя метод агрегации, и получаем окончательный рейтинг систем.

3.6.2. Эмпирическая оценка

Экспериментальный дизайн включает в себя четыре практических сценария. Эмпирическая оценка проводится на бенчмарках GLUE, SuperGLUE и VALUE в рамках сравнения результатов ранжирования с базовыми решениями: (i) средним арифметическим (σ^{am}), (ii) средним

геометрическим (σ^{gm}) и (iii) недостатком оптимальности (англ. *optimality gap*, σ^{og} [1]), — методом ранжирования, определяющим, насколько система отстает от минимального оптимального показателя качества $\gamma = 0,95$ (чем меньше, тем лучше).

Переранжирование бенчмарков

Описание практического сценария. В первом практическом сценарии рассматривается переранжирование бенчмарков в соответствии со скоринговыми и мажоритарными правилами и выбор победителя в соответствии со всеми предложенными правилами. Сравнение с базовыми решениями проводится путем вычисления (i) коэффициента согласия (англ. *agreement rate*, AR; в %), т.е. доли пересечения лучших и худших k систем между данным методом агрегации и σ^{am}, (ii) анализа корреляции τ Кендалла (τ) между полученными рейтингами систем [?], (iii) различающей способности (англ. *discriminative power*, DP), т.е. количеством эквивалентных альтернатив в рейтинге [16] и (iv) независимости от посторонних альтернатив (англ. *independence of irrelevant alternatives*, IIA), показывающей, как часто новая система в рейтинге влияет на окончательный рейтинг.

Ключевые результаты. Мы находим, что (i) методы агрегации согласны между собой относительно лучших и худших k систем при определенных значениях k, но предлагают отличающиеся друг от друга ранжирования систем, (ii) *правило Даудолла* и *правило Борда* чаще выбирают только одну лучшую альтернативу, в то время как *правило большинства* и *правило минимакса* выбирают значительно большее количество эквивалентных альтернатив, (iii) ранжирования, полученные *правилами Коупленда*, *минимакса* и *большинства*, менее прочих подвержены появлением новой системы в рейтинге, (iv) люди могут занимать лидирующие позиции на GLUE в соответствии с *правилами множества* и *Даудолла*, поскольку они все еще превосходят языковые модели на задачах определения логической и причинно-следствеенной связи [118; 129], (v) человек выбирается победителем на SuperGLUE *правилами большинства, Коупланда и Даудолла*, и, наконец, (vi) человек получает первое место в рейтинге систем VALUE в соответствии с *правилами Коупленда, минимакса и Кондорсе*, поскольку доминирует над системами при попарном сравнении.

Потенциальный победитель Кондорсе

Описание практического сценария. *Правило Кондорсе* объявляет победителем альтернативу, которая доминирует над всеми другими альтернативами при попарном сравнении [14]. Кроме того, устойчивость этого правила сложно нарушить, и оно легко интерпретируется. Во втором практическом сценарии решается следующая задача: найти такой вектор весов критериев, при котором интересующая альтернатива становится победителем Кондорсе, или определить, что такого вектора весов не существует.

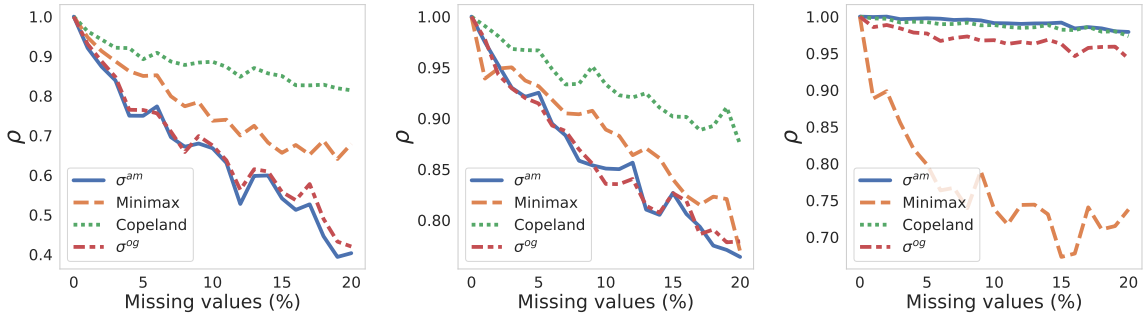

Рис. 6: Корреляция Спирмена (ρ) между топ-7 позициями в рейтингах систем GLUE, SuperGLUE и VALUE с опущенными и без опущенных значений метрик качества.

Ключевые результаты. Для 9, 10 и 3 систем на бенчмарках GLUE, SuperGLUE и VALUE соответственно можно найти победителя Кондорсе. Предложенный подход полезен для практических сценариев, в которых необходимо найти систему, которая доминирует над всеми другими при заданном наборе критериев оценки.

Оценка устойчивости к пропущенным значениям метрик качества

Описание практического сценария. В третьем практическом сценарии представлен более подробный анализ мажоритарных правил, отличительном свойством которых является возможность агрегации результатов с пропущенными значениями критериев оценки. Мы оцениваем устойчивость *правил Коупленда и Минимакса* и базовых решений σ^{am} и σ^{og} к пропущенным значениям метрик качества на рассматриваемых бенчмарках. Оценка проводится следующим образом. Первым шагом мы считаем ранжирования для каждого бенчмарка в исходном виде и используем их как золотой стандарт. Далее мы случайным образом заменяем N значений метрик качества на пустые значения и считаем ранжирования 7 лучших систем на полученных рейтингах систем. Мы считаем корреляцию Спирмена (ρ) между полученным ранжированием и золотым стандартом. Для базовых решений σ^{am} и σ^{og} мы используем медианные значения на соответствующих задачах. Эксперимент повторяется 100 раз, после чего результаты усредняются.

Ключевые результаты. На Рис. 6 представлены результаты данного эксперимента. σ^{am} и σ^{og} демонстрируют наименьшую устойчивость, в то время как *правило Коупленда* наиболее устойчиво на бенчмарках GLUE и SuperGLUE среди рассматриваемых методов. Однако *правило минимакса* наименее устойчиво на бенчмарке VALUE, что мы относим к его ограничению. Данное поведение характеризуется незначительной разницей в показателях качества систем на открытом рейтинге VALUE, в связи с чем *правило минимакса* присваивает системам похожий вес и чувствительно к любым пропущенным значениям метрик качества.

Ранжирование языковых моделей с учетом предпочтений пользователя

Описание практического сценария. Четвертый практический сценарий ориентирован на многокритериальный дизайн оценки языковых моделей с учетом предпочтения пользователя. В качестве предпочтений используется показатель качества на GLUE, вычислительная эффективность и предубеждения относительно социальных меньшинств. В данном эксперименте используются языковые модели BERT-base, RoBERTa-base, ALBERT-base, DeBERTa-base [43], DistilBERT-base [95], DistilRoBERTa-base [95] и GPT2-medium. Мы дообучаем модели на соответствующих задачах и оцениваем (i) показатели качества на валидационной выборке GLUE, (ii) вычислительную эффективность с помощью библиотеки Impact tracker [44] и (iii) показатели предубеждения относительно социальных меньшинств на наборах данных CrowS-Pairs [72], StereoSet [71] и Winobias [139].

Ключевые результаты. Мы излагаем ключевые результаты с использованием *правила Борда* и вектора веесов $(0.4, 0.3, 0.3)$, присвоенного для оценки качества на задачах GLUE, вычислительной эффективности и предубеждения относительно социальных меньшинств. Дистилированные языковые модели (DistilRoBERTa и DistilBERT) являются победителями, в то время как лидирующие на бенчмарке системы менее предпочтительны в связи с их неоптимальной вычислительной эффективностью и удовлетворительным показателем качества на наборах данных для оценки предубеждений. Сравнивая результаты со смежной работой [64], мы находим, что ранжирование относительно целевого качества не сохраняется, что объясняется тем, что в данной работе показателю качества на бенчмарке присваивается вес 0.5, который блокирует существенные изменения при повторном ранжировании.

3.6.3. Ретроспектива

Вопрос о том, каким образом агрегировать результаты эталонного тестирования, относится к каждому бенчмарку, предложенному в рамках данного диссертационного исследования. Vote'n'Rank обеспечивает возможность новой интерпретации *насыщенных* бенчмарков (англ. *saturated*), на которых улучшение качества становится незначительным или на которых системы достигли уровня человека. Проблема насыщения бенчмарков широко обсуждается в сообществе NLP, особенно в свете тенденций незначительного прироста качества большим количеством ресурсов [89; 86; 53]. Данная критика бесппорно важна для развития эталонного тестирования, однако она опирается на выводы, сделанные на основе использования среднего арифметического при агрегации результатов. Vote'n'Rank, напротив, позволяет сделать вывод о том, что — на момент публикации — люди все еще превосходят языковые модели в сценариях оценки, учитывающих предпочтения пользователя.

Сравнение систем с помощью Vote'n'Rank может быть затруднительным в связи с отсутствием золотого стандарта, особенно при отсутствующих значениях метрик качества (например, в

задачах поиска видео по текстовому запросу и генерации описания к видео[33]). Несмотря на это, мы вносим прикладной вклад в направление эталонного тестирования, предлагая альтернативные методы агрегации результатов оценки систем независимо от области МО. Наш фреймворк становится более актуальным в свете развивающейся парадигмы эталонного тестирования, которая нацелена на исчерпывающую оценку языковых моделей в сценариях, ориентированных на конечного пользователя [61].

4. Заключение

В заключительном разделе мы резюмируем основной вклад данного диссертационного исследования. Предложенные бенчмарки, кодовая база, открытые рейтинги систем, проекты по оценке людей на задачах NLU и другие материалы находятся в открытом доступе под лицензией Apache 2.0.

1. Предложены бенчмарки Russian SuperGLUE, RuCoLA и RuATD, которые стали стандартными ресурсами для измерения развития предобученных моделей для русского языка. В общей сложности бенчмарки включают в себя 11 задач NLU в различных формулировках, такие как машинное чтение, разрешение лексической многозначности, разрешение кореференции, определение логической и причинно-следственной связи, оценка приемлемости предложения, определение автора текста и распознавание сгенерированного текста. Каждый бенчмарк имеет открытый рейтинг систем для сравнения передовых языковых моделей относительно уровня человека.
2. Разработан новый метод автоматического распознавания сгенерированного текста, использующий три типа интерпретируемых признаков, извлеченных из графового представления матриц внимания языковой модели с помощью методов TDA. Данные признаки демонстрируют чувствительность к поверхностным и синтаксическим признакам текста. Разработанный метод превосходит по качеству существующие смежные решения на трех наборах данных, относящихся к разным предметным областям, и более устойчив к генеративным языковым моделям GPT-2, отсутствовавшим в обучающей выборке.
3. Предложен фреймворк Vote'n'Rank для ранжирования языковых моделей на бенчмарках, состоящих из множества задач, а так же при эталонном тестировании, учитывающем множество критериев оценки. Фреймворк включает в себя восемь интерпретируемых методов агрегации результатов по принципам теории многокритериального принятия решений. Vote'n'Rank позволяет решить проблемы, связанные с применением среднего арифметического при эталонном тестировании. Мы предлагаем рекомендации по использованию фреймворка на основе теоретических свойств методов агрегации результатов и сценариев применения в исследовательских и практических целях.

[33] value-benchmark.github.io/leaderboard.html

4. С использованием предложенных бенчмарков и инструментов эталонного тестирования проведена эмприческая оценка более 100 языковых моделей и их конфигураций в различных экспериментальных постановках. Мы излагаем ключевые результаты: (i) языковые модели значительно уступают людям на большинстве задач NLU, в то время как люди уступают языковым моделям в задаче распознавания сгенерированного текста, (ii) языковые модели в меньшей степени чувствительны к морфологическим и семантическим нарушениям в неприемлемых предложениях, но демонстрируют хорошую обобщающую способность при оценке приемлемости сгенерированных предложений, (iii) мультиязычные языковые модели достигают низкого качества при переносе знания грамматики между русским, английским и итальянским языками, (iv) языковые модели, занимающие лидирующие позиции на стандартных бенчмарках, могут быть менее предпочтительны в многокритериальных сценариях оценки, дополнительно учитывающих вычислительную эффективность и предубеждения относительно социальных меньшинств, и (v) люди могут занимать лидирующие позиции на насыщенных бенчмарках в соответствии с методами агрегации результатов Vote'n'Rank, поскольку они все еще превосходят языковые модели по качеству решения определенных задач NLU.

Список литературы

[1] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep Reinforcement Learning at the Edge of the Statistical Precipice. *Advances in Neural Information Processing Systems*, 34, 2021.

[2] Mark Aizerman and Fuad Aleskerov. *Theory of Choice*, volume 38. North Holland, 1995.

[3] Fuad Aleskerov, Vyacheslav V Chistyakov, and Valery Kalyagin. The threshold aggregation. *Economics Letters*, 107(2):261--262, 2010.

[4] Amazon. Amazon Customer Reviews Dataset. `https://s3.amazonaws.com/amazon-reviews-pds/readme.html`, 2019.

[5] Alexandra Antonova and Alexey Misyurev. Building a web-based parallel corpus and filtering out machine-translated text. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 136--144, Portland, Oregon, June 2011. Association for Computational Linguistics.

[6] Kenneth J Arrow. Social Choice and Individual Values. In *Social Choice and Individual Values*. Yale university press, 2012.

[7] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623--4637, Online, July 2020. Association for Computational Linguistics.

[8] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597--610, 2019.

[9] Sameer Badaskar, Sachin Agarwal, and Shilpa Arora. Identifying real or fake articles: Towards better language modeling. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*, 2008.

[10] Shreyan Bakshi, Soumya Batra, Peyman Heidari, Ankit Arun, Shashank Jain, and Michael White. Structure-to-text generation with self-training, acceptability classifiers and context-conditioning for the GEM shared task. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 136--147, Online, August 2021. Association for Computational Linguistics.

[11] Soumya Batra, Shashank Jain, Peyman Heidari, Ankit Arun, Catharine Youngs, Xintong Li, Pinar Donmez, Shawn Mei, Shiunzu Kuo, Vikas Bhardwaj, Anuj Kumar, and Michael White. Building adaptive acceptability classifiers for neural NLG. In *Proceedings of the 2021 Conference*

on Empirical Methods in Natural Language Processing, pages 682--697, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[12] Anya Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. The ReproGen Shared Task on Reproducibility of Human Evaluations in NLG: Overview and Results. In *The 14th International Conference on Natural Language Generation*, 2021.

[13] Emily Bender. The BenderRule: On Naming the Languages We Study and Why It Matters. *The Gradient*, 2019.

[14] Duncan Black et al. The theory of committees and elections. 1958.

[15] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*, 2021.

[16] Felix Brandt and Hans Georg Seedig. On the Discriminative Power of Tournament Solutions. In *Operations Research Proceedings 2014*, pages 53--58. Springer, 2016.

[17] Franck Burlot, Yves Scherrer, Vinit Ravishankar, Ondřej Bojar, Stig-Arne Grönroos, Maarit Koponen, Tommi Nieminen, and François Yvon. The WMT'18 morpheval test suites for English-Czech, English-German, English-Finnish and Turkish-English. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 546--560, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[18] Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhreva, and Marat Zaynutdinov. DeepPavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122--127, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[19] Frédéric Chazal and Bertrand Michel. An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. *Frontiers in artificial intelligence*, 4:667963, 2021.

[20] Daniil Cherniavskii, Eduard Tulchinskii, Vladislav Mikhailov, Irina Proskurina, Laida Kushnareva, Ekaterina Artemova, Serguei Barannikov, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. Acceptability judgements via examining the topology of attention maps. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 88--107, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[21] Noam Chomsky. *Aspects of the Theory of Syntax*. MIT Press, 1965.

[22] Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. Rethinking Embedding Coupling in Pre-trained Language Models. In *International Conference on Learning Representations*, 2020.

[23] Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454--470, 2020.

[24] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276--286, Florence, Italy, August 2019. Association for Computational Linguistics.

[25] James Clough, Nicholas Byrne, Ilkay Oksuz, Veronika A Zimmer, Julia A Schnabel, and Andrew King. A Topological Loss Function for Deep Learning-based Image Segmentation Using Persistent Homology. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[26] Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stephan Clemencon. What are the Best Systems? New Perspectives on NLP Benchmarking. In *Advances in Neural Information Processing Systems*, 2022.

[27] Pierre Colombo, Guillaume Staerman, Chloé Clavel, and Pablo Piantanida. Automatic text evaluation through the lens of Wasserstein barycenters. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10450--10466, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[28] Pierre Jean A Colombo, Chloé Clavel, and Pablo Piantanida. InfoLM: A New Metric to Evaluate Summarization & Data2Text Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10554--10562, 2022.

[29] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440--8451, Online, July 2020. Association for Computational Linguistics.

[30] Jurafsky Daniel, Martin James H, et al. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2007.

[31] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20--28, 1979.

[32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171--4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[33] Liam Dugan, Daphne Ippolito, Arun Kirubarajan, and Chris Callison-Burch. RoFT: A tool for evaluating human detection of machine-generated text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 189--196, Online, October 2020. Association for Computational Linguistics.

[34] Pavel Efimov, Andrey Chertok, Leonid Boytsov, and Pavel Braslavski. SberQuAD--Russian Reading Comprehension Dataset: Description and Analysis. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22--25, 2020, Proceedings 11*, pages 3--15. Springer, 2020.

[35] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. Beyond English-centric Multilingual Machine Translation. *Journal of Machine Learning Research*, 22(107):1--48, 2021.

[36] Alena Fenogenova. Russian paraphrasers: Paraphrase with transformers. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 11--19, Kiyv, Ukraine, April 2021. Association for Computational Linguistics.

[37] Alena Fenogenova, Vladislav Mikhailov, and Denis Shevelev. Read and reason with MuSeRC and RuCoS: Datasets for machine reading comprehension for Russian. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6481--6497, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[38] Alena Fenogenova, Tatiana Shavrina, Alexandr Kukushkin, Maria Tikhonova, Anton Emelyanov, Valentin Malykh, Vladislav Mikhailov, Denis Shevelev, and Ekaterina Artemova. Russian SuperGLUE 1.1: Revising the Lessons not Learned by Russian NLP-models. 2021.

[39] Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111--116, Florence, Italy, July 2019. Association for Computational Linguistics.

[40] Ilya Gusev. Dataset for automatic summarization of russian news. In *Artificial Intelligence and Natural Language*, pages 122--134. Springer International Publishing, 2020.

[41] Mareike Hartmann, Miryam de Lhoneux, Daniel Hershcovich, Yova Kementchedjhieva, Lukas Nielsen, Chen Qiu, and Anders Søgaard. A multilingual benchmark for probing negation-awareness with minimal pairs. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 244--257, Online, November 2021. Association for Computational Linguistics.

[42] Jie He, Bo Peng, Yi Liao, Qun Liu, and Deyi Xiong. TGEA: An error-annotated dataset and benchmark tasks for TextGeneration from pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6012--6025, Online, August 2021. Association for Computational Linguistics.

[43] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*, 2020.

[44] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. *Journal of Machine Learning Research*, 21(248):1--43, 2020.

[45] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation. In *International Conference on Machine Learning*, pages 4411--4421. PMLR, 2020.

[46] Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808--1822, Online, July 2020. Association for Computational Linguistics.

[47] Vitaly Ivanin, Ekaterina Artemova, Tatiana Batura, Vladimir Ivanov, Veronika Sarkisyan, Elena Tutubalina, and Ivan Smurov. Rurebus-2020 shared task: Russian relation extraction for business. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialogue"]*, Moscow, Russia, 2020.

[48] Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296--2309, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[49] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282--6293, Online, July 2020. Association for Computational Linguistics.

[50] Marzena Karpinska, Nader Akoury, and Mohit Iyyer. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265--1285, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[51] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252--262, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[52] Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, Mozhdeh Gheini, Arman Kabiri, Rabeeh Karimi Mahabagdi, Omid Memarrast, Ahmadreza Mosallanezhad, Erfan Noury, Shahab Raji, Mohammad Sadegh Rasooli, Sepideh Sadeghi, Erfan Sadeqi Azer, Niloofar Safi Samghabadi, Mahsa Shafaei, Saber Sheybani, Ali Tazarv, and Yadollah Yaghoobzadeh. ParsiNLU: A suite of language understanding challenges for Persian. *Transactions of the Association for Computational Linguistics*, 9:1147--1162, 2021.

[53] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110--4124, Online, June 2021. Association for Computational Linguistics.

[54] Samuel Kounev, Klaus-Dieter Lange, and Jóakim von Kistowski. *Systems Benchmarking: For Scientists and Engineers*, volume 1. Springer, 2020.

[55] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957--2966, Marseille, France, June 2022. European Language Resources Association.

[56] Laida Kushnareva, Daniil Cherniavskii, Vladislav Mikhailov, Ekaterina Artemova, Serguei Barannikov, Alexander Bernstein, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. Artificial text detection via examining the topology of attention maps. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 635--649, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[57] Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. How furiously can colorless green ideas sleep? sentence acceptability in context. *Transactions of the Association for Computational Linguistics*, 8:296--310, 2020.

[58] Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483--4499, Online, November 2020. Association for Computational Linguistics.

[59] Byron C Lewis and Albert E Crews. The Evolution of Benchmarking as a Computer Performance Evaluation Technique. *MIS Quarterly*, pages 7--16, 1985.

[60] Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang Wang, et al. VALUE: A Multi-Task Benchmark for Video-and-Language Understanding Evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[61] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic Evaluation of Language Models. *arXiv preprint arXiv:2211.09110*, 2022.

[62] Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008--6018, Online, November 2020. Association for Computational Linguistics.

[63] Ekaterina Lutikova. K voprosu o kategorial'nom statuse imennykh grup v russkom yazyke. *Moscow University Philology Bulletin*, 2010.

[64] Zhiyi Ma, Kawin Ethayarajh, Tristan Thrush, Somya Jain, Ledell Wu, Robin Jia, Christopher Potts, Adina Williams, and Douwe Kiela. Dynaboard: An Evaluation-As-A-Service Platform for Holistic Next-Generation Benchmarking. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10351--10367. Curran Associates, Inc., 2021.

[65] Brian W. Matthews. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochimica et biophysica acta*, 405 2:442--51, 1975.

[66] Kris McGuffie and Alex Newhouse. The Radicalization Risks of GPT-3 and Advanced Neural Language Models. *arXiv preprint arXiv:2009.06807*, 2020.

[67] Michail Melnichenko and Natalia Tyshkevich. Prozhito from Manuscript to Corpus. *ISTORIYA*, 8(7 (61)), 2017.

[68] Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. RuCoLA: Russian corpus of linguistic acceptability. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5207--5227, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[69] Swaroop Mishra and Anjana Arunkumar. How Robust are Model Rankings: A Leaderboard Customization Approach for Equitable Evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13561--13569, 2021.

[70] Olga Mitrenina, Evgeniya Romanova, and Natalia Slioussar. *Vvedeniye v generativnuyu grammatiku*. Limited Liability Company "LIBROCOM", 2017.

[71] Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356--5371, Online, August 2021. Association for Computational Linguistics.

[72] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953--1967, Online, November 2020. Association for Computational Linguistics.

[73] Christina Nießl, Moritz Herrmann, Chiara Wiedemann, Giuseppe Casalicchio, and Anne-Laure Boulesteix. Over-optimism in Benchmark Studies and the Multiplicity of Design and Analysis Options when Interpreting Their Results. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(2):e1441, 2022.

[74] Hannu Nurmi. Voting procedures: A summary analysis. *British Journal of Political Science*, 13(2):181--208, 1983.

[75] Elena Paducheva. *Dinamicheskiye modeli v semantike leksiki*. Languages of Slavonic culture, 2004.

[76] Elena Paducheva. *Semanticheskiye issledovaniya: Semantika vremeni i vida v russkom yazyke*. Languages of Slavonic culture, second edition, 2010.

[77] Elena Paducheva. *Russkoye otritsatel'noye predlozheniye*. Languages of Slavonic culture, 2013.

[78] Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, et al. KLUE: Korean Language

Understanding Evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[79] Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alexia Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R. Bowman. Does putting a linguist in the loop improve NLU data collection? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4886--4901, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[80] Adam Pauls and Dan Klein. Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 959--968, 2012.

[81] Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74--84, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[82] Irina Proskurina, Irina Piontkovskaya, and Ekaterina Artemova. Can BERT Eat RuCoLA? Topological Data Analysis to Explain. *arXiv preprint arXiv:2304.01680*, 2023.

[83] Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529--535, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[84] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language Models are Unsupervised Multitask Learners. 2019.

[85] Mark Rofin, Vladislav Mikhailov, Mikhail Florinsky, Andrey Kravchenko, Tatiana Shavrina, Elena Tutubalina, Daniel Karabekyan, and Ekaterina Artemova. Vote'n'rank: Revision of benchmarking with social choice theory. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 670--686, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[86] Anna Rogers. How the Transformers Broke NLP Leaderboards. `https://hackingsemantics.xyz/2019/leaderboards`, 2019.

[87] Anna Rogers, Matt Gardner, and Isabelle Augenstein. QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension. *ACM Computing Surveys*, 55(10):1--45, 2023.

[88] Sebastian Ruder. Why You Should Do NLP Beyond English. http://ruder.io/nlp-beyond-english, 2020.

[89] Sebastian Ruder. Challenges and Opportunities in NLP Benchmarking. http://ruder.io/nlp-benchmarking, 2021.

[90] Sebastian Ruder. The State of Multilingual AI. http://ruder.io/state-of-multilingual-ai/, 2022.

[91] Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215--10245, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[92] Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. KLEJ: Comprehensive benchmark for Polish language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1191--1201, Online, July 2020. Association for Computational Linguistics.

[93] Andrey Sakhovskiy, Alexandra Izhevskaya, Alena Pestova, Elena Tutubalina, Valentin Malykh, Ivan Smurov, and Ekaterina Artemova. RuSimpleSentEval-2021 Shared Task: Evaluating Sentence Simplification for Russian. In *Proceedings of the International Conference "Dialogue"*, pages 607--617, 2021.

[94] Gerard Salton and C. S. Yang. On the specification of term values in automatic indexing. *Journal of Documentation*, 29:351--372, 1973.

[95] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[96] Ketki Savle, Wlodek Zadrozny, and Minwoo Lee. Topological data analysis for discourse semantics? In *Proceedings of the 13th International Conference on Computational Semantics - Student Papers*, pages 34--43, Gothenburg, Sweden, May 2019. Association for Computational Linguistics.

[97] Carson T. Schütze. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press, 1996.

[98] Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351--1361, Online, April 2021. Association for Computational Linguistics.

[99] Haitham Seelawi, Ibraheem Tuffaha, Mahmoud Gzawi, Wael Farhan, Bashar Talafha, Riham Badawi, Zyad Sober, Oday Al-Dweik, Abed Alhakim Freihat, and Hussein Al-Natsheh. ALUE: Arabic language understanding evaluation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 173--184, Kyiv, Ukraine (Virtual), April 2021. Association for Computational Linguistics.

[100] Olga Seliverstova. *Trudy po semantike*. Languages of Slavonic culture, 2004.

[101] Tatiana Shamardina, Vladislav Mikhailov, Daniil Chernianskii, Alena Fenogenova, Marat Saidov, Anastasiya Valeeva, Tatiana Shavrina, Ivan Smurov, Elena Tutubalina, and Ekaterina Artemova. Findings of the RuATD Shared Task 2022 on Artificial Text Detection in Russian. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2022"*, 2022.

[102] Tatiana Shavrina, Anton Emelyanov, Alena Fenogenova, Vadim Fomin, Vladislav Mikhailov, Andrey Evlampiev, Valentin Malykh, Vladimir Larin, Alex Natekin, Aleksandr Vatulin, Peter Romov, Daniil Anastasiev, Nikolai Zinov, and Andrey Chertok. Humans keep it one hundred: an overview of AI journey. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2276--2284, Marseille, France, May 2020. European Language Resources Association.

[103] Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717--4726, Online, November 2020. Association for Computational Linguistics.

[104] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. Release Strategies and the Social Impacts of Language Models, 2019.

[105] Anirudh Som, Hongjun Choi, Karthikeyan Natesan Ramamurthy, and Matthew P. Buman. Pi-net: A Deep Learning Approach to Extract Topological Persistence Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[106] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning, 2020.

[107] Yakov Testelets. *Vvedeniye v obschiy sintaksis*. Russian State University for the Humanities, 2001.

[108] Jörg Tiedemann and Santhosh Thottingal. OPUS-MT -- building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine*

Translation, pages 479--480, Lisboa, Portugal, November 2020. European Association for Machine Translation.

[109] Daniela Trotta, Raffaele Guarasci, Elisa Leonardelli, and Sara Tonelli. Monolingual and cross-lingual acceptability judgments with the Italian CoLA corpus. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2929--2940, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[110] Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Serguei Barannikov, Irina Piontkovskaya, Sergey Nikolenko, and Evgeny Burnaev. Topological Data Analysis for Speech Processing. *arXiv preprint arXiv:2211.17223*, 2022.

[111] Alan M Turing and J Haugeland. Computing machinery and intelligence. *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*, pages 29--56, 1950.

[112] Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384--8395, Online, November 2020. Association for Computational Linguistics.

[113] Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2001--2016, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[114] Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. BasqueGLUE: A natural language understanding benchmark for Basque. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1603--1612, Marseille, France, June 2022. European Language Resources Association.

[115] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. *Advances in Neural Information Processing Systems*, 30, 2017.

[116] Elena Volodina, Yousuf Ali Mohammed, and Julia Klezl. DaLAJ -- a dataset for linguistic acceptability judgments for Swedish. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 28--37, Online, May 2021. LiU Electronic Press.

[117] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *Advances in Neural Information Processing Systems*, 32, 2019.

[118] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding.

In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353--355, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[119] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *International Conference on Learning Representations*, 2019.

[120] Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. Adversarial GLUE: A Multi-Task Benchmark for Robustness Evaluation of Language Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[121] W Wang, N Yang, F Wei, B Chang, and M Zhou. R-NET: Machine Reading Comprehension with Self-matching Networks. *Natural Lang. Comput. Group, Microsoft Res. Asia, Beijing, China, Tech. Rep*, 5, 2017.

[122] Alex Warstadt and Samuel R Bowman. Linguistic Analysis of Pretrained Sentence Encoders with Acceptability Judgments. *arXiv preprint arXiv:1901.03438*, 2019.

[123] Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377--392, 2020.

[124] Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625--641, 2019.

[125] Geoffrey I Webb. MultiBoosting: A Technique for Combining Boosting and Wagging. *Machine learning*, 40(2):159--196, 2000.

[126] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and Social Risks of Harm from Language Models, 2021.

[127] Mattthew E Werenski, Ruijie Jiang, Abiy Tasissa, Shuchin Aeron, and James M Murphy. Measure Estimation in the Barycentric Coding Model. In *International Conference on Machine Learning*, pages 23781--23803. PMLR, 2022.

[128] Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational*

Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 843--857, Suzhou, China, December 2020. Association for Computational Linguistics.

[129] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112--1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[130] Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784--2790, Online, April 2021. Association for Computational Linguistics.

[131] Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762--4772, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[132] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483--498, Online, June 2021. Association for Computational Linguistics.

[133] Fan Yin, Quanyu Long, Tao Meng, and Kai-Wei Chang. On the robustness of language encoders against grammatical errors. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3386--3403, Online, July 2020. Association for Computational Linguistics.

[134] Kuratov Yuri and Arkhipov Mikhail. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019"*, 2019.

[135] Aleš Žagar and Marko Robnik-Šikonja. Slovene SuperGLUE benchmark: Translation and evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2058--2065, Marseille, France, June 2022. European Language Resources Association.

[136] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *Advances in neural information processing systems*, 32, 2019.

[137] Changchang Zeng, Shaobo Li, Qin Li, Jie Hu, and Jianjun Hu. A Survey on Machine Reading Comprehension -- Tasks, Evaluation Metrics and Benchmark Datasets. *Applied Sciences*, 10(21):7640, 2020.

[138] Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. ReCoRD: Bridging the Gap Between Human and Machine Commonsense Reading Comprehension. *arXiv preprint arXiv:1810.12885*, 2018.

[139] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15--20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[140] Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393--1404, Online, August 2021. Association for Computational Linguistics.

[141] Xiyou Zhou, Zhiyu Chen, Xiaoyong Jin, and William Yang Wang. HULK: An energy efficiency benchmark platform for responsible natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 329--336, Online, April 2021. Association for Computational Linguistics.