О методах компьютерной лингвистики в оценке систем искусственного интеллекта

Татьяна Олеговна Шаврина

Национальный исследовательский университет «Высшая школа экономики», Москва, Россия; Институт искусственного интеллекта, Москва, Россия; Институт проблем передачи информации им. А. А. Харкевича РАН, Москва, Россия; ООО «СберДевайсы», Москва, Россия; shavrina@airi.net

Аннотация: В статье рассматриваются актуальные исследования в области прикладной лингвистики, посвященные оценке систем искусственного интеллекта (ИИ). В качестве основного инструмента для оценки уровня интеллектуальности систем выступают языковые тесты. Они являются самым доступным способом обучения систем ИИ и одновременно обладают высокой вариативностью, необходимой для формулировки интеллектуальных задач. Приводится обзор актуальной методологии обучения и тестирования интеллектуальных систем, рассматриваются золотые стандарты текстовых задач (бенчмарки) в методологии General Language Understanding Evaluation (GLUE). Обсуждаются теоретические основы и конкретные реализации теста для ИИ-систем «Russian SuperGLUE». Дальнейшее сближение практик машинного обучения и науки о языке способно заполнить лакуны как в оценке ИИ-систем, так и в методах их эффективного обучения.

Ключевые слова: автоматический анализ текста, искусственный интеллект, компьютерная лингвистика, машинное обучение, обзор

Благодарности: Работа выполнена при поддержке гранта Министерства науки и высшего образования № 075-15-2020-793.

Для цитирования: Шаврина Т. О. О методах компьютерной лингвистики в оценке систем искусственного интеллекта. *Вопросы языкознания*, 2021, 6: 117–138.

DOI: 10.31857/0373-658X.2021.6.117-138

Methods of computational linguistics in the evaluation of artificial intelligence systems

Tatiana O. Shavrina

HSE University, Moscow, Russia; Artificial Intelligence Research Institute, Moscow, Russia; Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia; SberDevices LLC, Moscow, Russia; shavrina@airi.net

Abstract: The article discusses current research in the field of applied linguistics dedicated to the evaluation of artificial intelligence (AI) systems. Linguistic tests are used as the principal tool for evaluating the level of intelligence of such systems, being the most affordable way of training AI systems and, at the same time, having high variability necessary for the formulation of intellectual tasks. This paper provides an overview of current methodology for training and testing AI systems and describes the gold standards of textual tasks (benchmarks) in the General Language Understanding Evaluation (GLUE) methodology. We also present an overview of how the theoretical apparatus and practices of linguistics are used to create a Russian-language test for examining the abilities of AI systems, the Russian

SuperGLUE. Further convergence of machine learning and linguistic methods can fill gaps both in the practice of evaluating AI systems and in their effective training.
Keywords: artificial intelligence, computational linguistics, machine learning, natural language processing, survey
Acknowledgements: The work was supported by the Ministry of Science and Higher Education of Russia, grant No. 075-15-2020-793.
For citation: Shavrina T. O. Methods of computational linguistics in the evaluation of artificial intelligence systems. *Voprosy Jazykoznanija*, 2021, 6: 117–138.
DOI: 10.31857/0373-658X.2021.6.117-138

"Human knowledge is expressed in language.
So computational linguistics is very important."
Mark Steedman, ACL Presidential Address (2007)

1. Введение: подходы к оценке искусственного интеллекта

Язык, уникальная человеческая способность, использующая самые разные отделы мозга и включенная в само определение искусственного интеллекта (ИИ), пока еще не моделируется достаточно успешно компьютерными системами. Для отслеживания прогресса в этой области с 1990-х гг. проводятся тестирования; все они так или иначе основаны на тесте Тьюринга [Turing 1950]. Этот тест приравнивает языковую способность к интеллектуальности: участники-судьи обмениваются с тестируемой системой сообщениями вслепую, а система должна убедить в собственной «человечности» не менее определенного процента судей.

Такой подход, с одной стороны, подвержен необъективности за счет человеческого фактора; с другой стороны, он может быть бесконечно масштабируем: сообщения судей могут включать различные вопросы, задания, просьбы, и т. д., что расширяет возможности для более объективной оценки. Но можно ли определить интеллектуальность количественно?

Определение искусственного интеллекта (и его наличие вообще) во многом опирается на ключевые признаки, которыми должна обладать система. Выделяют сильный и слабый ИИ. Для т. н. **сильного ИИ** такими признаками являются способность к принятию решений, использованию стратегий, решению головоломок и действиям в условиях неопределенности; а также обладание знаниями и способность к их представлению; общие навыки планирования, самообучения, общения на естественном языке и объединение всех этих способностей воедино для достижения общих целей (список основан на [Nilsson 1998; Poole et al. 1998; Russell, Norvig 2003; Luger, Stubblefield 2004]).

Слабый ИИ способен решать узконаправленные задачи и перечисленными выше признаками обладать не обязан: слабому ИИ соответствует актуальное состояние технического прогресса — такие успешно решаемые задачи, как реферирование и перевод[1] текстов, вождение автомобиля[2], игра в шахматы и го[3], являются хорошим примером реализации слабого ИИ. К слабым ИИ-системам относится, например, ПО для автоматического решения единого государственного экзамена по русскому языку, сочетающее непосредственные текстовые источники знаний (тексты учебников), статистические модели ранжирования ответов, несколько моделей для расстановки пунктуации, нейросетевую систему проверки орфографии, систему правил для решения заданий на понимание текста и нейросетевую

[1] https://www.microsoft.com/en-us/research/uploads/prod/2018/03/final-achieving-human.pdf

[2] https://en.wikipedia.org/wiki/Self-driving_car

[3] https://en.wikipedia.org/wiki/AlphaGo

модель для генерации текста сочинения [Shavrina et al. 2020a]. Инженеру или составителю ЕГЭ при работе с системой станет ясно, что она не является в полной мере интеллектуальной, так как лишь использует фиксированный набор правил и фактов, хотя и может продемонстрировать определенные, вполне удовлетворительные, результаты в рамках поставленной задачи. Ни каждая из ее составляющих в отдельности, ни их совокупность не обладают знанием о русском языке, однако в целом она демонстрирует уровень, достаточный для имитации успешного выполнения экзаменационных заданий.

Само понятие сильного ИИ, введенное Дж. Сёрлем в 1980 г. [Searle 1980], изначально отделяет сильный интеллект от слабого по наличию самосознания и восприятия; при этом слабый ИИ все равно считается полезным методологически [Frankish, Ramsey (eds.) 2014] — для проверки гипотез о сознании, — обладая отдельными способностями, но не являясь антропоморфной системой.

Помимо разделения на сильный и слабый ИИ, следует упомянуть разделение на широкий и узкий ИИ по количеству освоенных областей знания[4]. Распространено также разделение интеллектуальных систем на **унимодальные**, работающие с одной модальностью данных (текст, устная речь, изображения, видео, пространственная информация и т. д.), и **мультимодальные** [Baltrušaitis et al. 2018].

Фундаментальным в концепции Сёрля является понятие **общего ИИ** (Artificial General Intelligence, AGI), который определяется одновременно как

— сильный — умеющий не только выполнять фиксированный набор задач, но и учиться новым навыкам и ставить цели, планировать;
— широкий — умеющий делать обобщения в разных тематических сферах;
— мультимодальный — принимающий во внимание и текст, и изображения, и звуковую информацию и т. д.

Способ достижения AGI при этом может быть любым: «Соответствующим образом запрограммированный компьютер с нужными входами и выходами и будет разумом в том смысле, в котором человеческий разум — это разум» [Searle 1980: 417].

Как моделирующая наука лингвистика обладает широким спектром методов, которые наследуются другими дисциплинами, связанными с обработкой языковых данных. Оценка успешности моделирования отдельных элементов языковой системы и механизмов мышления, выражаемых с помощью языка, исключительно важна для систем искусственного интеллекта и для отслеживания их развития.

В отличие от человека, для которого устная речь по отношению к письменной первична, системам искусственного интеллекта приходится работать с текстовой информацией как с первичной, а распознавание текста на слух и озвучивание являются инженерным дополнением. Большинство современных ИИ-систем в качестве самого доступного, но опосредованного результата мышления человека получают для обучения языку / языкам корпуса текстов. На основании некоторой статистики слов[5] этих корпусов система должна приобрести аналогичные человеческим навыки владения языком, а именно составить словарь и обучиться словоизменению, синтаксису, базовым представлениям об упомянутых в текстах объектах, их свойствах и взаимодействии.

Высокоуровневая оценка и естественного, и искусственного интеллекта теснее связана с лингвистическими проблемами, чем с какими бы то ни было иными. Действительно, любая интеллектуальная задача может быть сформулирована с помощью естественного

[4] Узкий ИИ оперирует лишь в рамках небольшого ограниченного набора предметных областей, тогда как широкий является мультидоменным, использует полученные в одной области знания при работе с другой и строит на них совместные обобщения. В некоторых работах узкий ИИ терминологически приравнивается к слабому, т. е. решающему одну задачу в рамках одного домена: https://io9.gizmodo.com/how-much-longer-before-our-first-ai-catastrophe-464043243.

[5] Частоты n-грамм, частоты встречаемости слов в одном контексте и т. д.

языка — будь то детская загадка или задание из учебника по высшей математике. Описание задачи будет выражено в рамках языковой системы, включающей более и менее формальные подсистемы. Любую задачу, таким образом, можно представить как систему перехода из *X* в *y*, где *X* — текст задачи, а *y* — текст ответа в рамках этой же языковой системы [Raffel et al. 2019]. Составление таких задач может наследовать из лингвистики многие критерии оценки правильности моделирования:

— грамматические критерии:
  ○ успешность преодоления синтаксической и морфологической неоднозначности, разрешения анафоры (если система справляется с этим так же, как человек, то это рассматривается как успех моделирования);

— семантические критерии:
  ○ успешность снятия семантической неоднозначности, в том числе разрешения омонимии и полисемии в контексте (некоторая доля ошибок допустима у человека, но у искусственных систем она приводит к негативному восприятию собеседниками);
  ○ успешность речевых актов и соблюдение коммуникативных норм (некоторая доля неуспешных речевых актов, безусловно, есть и у человека, однако система, имеющая низкий процент удач, не сможет содержательно пообщаться, выполнить команды);
  ○ способность к корректной интерпретации высказываний, включающих кванторы, фактивность, предикаты различных импликативных типов, метафоры, некомпозициональность и т. д.;

— психолингвистические критерии:
  ○ подверженность прайммингу и различным психолингвистическим стимулам в рамках повествования, описания известной ситуации (актуализированность в сознании определенных понятий у человека может провоцировать ошибки; у машин таким триггером ошибки могут быть смещения в обучающем корпусе);
  ○ степень удобочитаемости (readability) для машин и человека (выразить один и тот же смысл человек может по-разному в различных обстоятельствах; система все равно должна корректно проинтерпретировать высказывание);

— типологические критерии:
  ○ успешность моделирования лексических закономерностей в разных языках и их влияние на различные навыки ИИ-систем, например проведение причинно-следственных связей. Отдельно оценивается успешность моделирования типичных для разных языков ассоциативных связей, адекватность употребления устойчивых выражений, упоминания культурных реалий — все на разных языках.

Лингвистические знания и методы в ИИ применяются исключительно широко: лишь один обзор того, как используются знания о различных уровнях и явлениях языка в разметке данных при решении узких задач машинного обучения, занял бы не одну сотню страниц. В фокусе внимания этой работы — использование лингвистических методов именно при оценке интеллектуальности систем.

Следующий, второй раздел посвящен доминирующему подходу в обсуждаемой области — системам оценки интеллекта, базирующимся на тесте Тьюринга. Рассматриваются их различные вариации для англоязычных и русскоязычных систем, а также для все более многочисленных многоязычных систем. Системы оценки интеллекта индифферентны по отношению к методам решения представленных задач: считается равно уместным решить задачу с помощью правил, с помощью статистического подхода, нейросетей, любой биологически вдохновленной системы, главное — доказать работоспособность решения. Методики тестирования усложняются, сталкиваясь с практическими проблемами,

например утечкой «золотых» ответов или иными ситуациями, в результате которых появляется возможность получить высокий балл за ответ без разработки решения, — подобные проблемы представлены в разделе 2.4. Заключает работу итоговый раздел 3.

2. Комплексные методы оценки ИИ, использующие языковые данные

Как отмечает Франсуа Шолле в фундаментальном обзоре истории оценки ИИ-систем [Chollet 2019: 3], цель развития ИИ — в создании машин с интеллектом, который сопоставим с интеллектом людей: такая цель была сформулирована в начале 50-х годов XX в., и с тех пор эта формулировка сохраняется.

На уровне практической разработки слабых ИИ-систем определения интеллекта до недавнего времени не требовалось, как не требовалось и общих критериев, объединяющих метрики качества решения сразу нескольких задач. Но увеличение количества публикаций и разработок в сфере машинного обучения[6] привело к необходимости методологической проработки стандартов описания моделей и всех этапов эксперимента, включая сбор и предварительную обработку данных для обучения, воспроизводимость результатов и условия тестирования[7], а главное — к необходимости создания общих измеримых критериев оценки интеллектуальности.

Чтобы эффективно повышать интеллектуальность искусственных систем, нам требуется не только ясное определение интеллекта, но и умение оценивать его. Это нужно для корректного сравнения двух систем между собой или для сопоставления системы с человеком.

Лингвистика обеспечивает неотъемлемую часть этой оценки, выступая в качестве источника тестов и методологических практик.

Предлагаемые ею тесты, тем не менее, за последние 10 лет подвергались существенной адаптации, переработке, усложнению. Процесс их переработки все еще не завершен, так как последние прикладные работы в моделировании языка достигли значительных успехов и формально справляются с тестированием. Стоит отметить, что успехи последних пяти лет были достигнуты за счет одновременного воздействия двух факторов:

1) использования новых нейросетей на основе encoder-decoder и механизма внимания;
2) увеличения обучающих корпусов и количества обучаемых параметров (коэффициентов) у нейросети: «чем больше, тем лучше».

Приведем доминирующие принципы работы этих факторов:

— Языковая модель — набор вероятностей встречаемости слов, букв и фраз в корпусе. Языковая модель может быть основана на классических методах теории вероятности (например, на цепях Маркова). С помощью языковых моделей можно оценивать вероятность употребления в тексте той или иной фразы («зеленые идеи яростно спят» — почти невозможная фраза, так как вероятность встретить такие словоформы рядом очень низка), а также предложить наиболее вероятное продолжение заданному началу фразы: если «шел проливной…», то далее будет «дождь» (вероятность 90 %), «ливень» (5 %), все остальные варианты — 5 %. Языковые модели сравнивают между собой по способности оценивать вероятность / невероятность новых для них

[6] https://www.technologyreview.com/2019/01/25/1436/we-analyzed-16625-papers-to-figure-out-where-ai-is-headed-next/

[7] Стандарты воспроизводимости конференции «Conference on Neural Information Processing Systems 2020»: https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist-v2.0.pdf.

фраз на т. н. «золотых корпусах»[8]. Самым используемым среди этих корпусов является Penn Treebank[9].

— Языковая модель на основе нейросети encoder-decoder и механизма внимания обладает несколькими нововведениями, которые позволяют таким моделям занимать лидирующие позиции в рейтингах моделирования языка. Во время обучения нейросеть проходит по всем текстам корпуса, удерживая в рабочей памяти до нескольких тысяч предыдущих слов. Нейросеть учится выполнять лишь одну операцию — по словам в ее памяти предсказать следующие слова: «шел проливной …». Механизм внимания во время такого обучения подбирает веса «важности» прочитанных слов в памяти, измеряя их влияние на продолжение текста. В процессе чтения текста модель штрафует сама себя за неправильное предсказание, корректируя веса механизма внимания. Таким образом нейросеть обучается правильно определять ключевые мысли в тексте, наибольшим образом влияющие на дальнейшее развитие мысли и дискурсивную структуру текста.

— Большинство ИИ-систем, показывающих лучшие результаты в существующих системах оценки интеллекта, основаны именно на таком механистическом подходе. Различные вариации механизма внимания, способов чтения текста и объема памяти составляют ряд нейросетей, называемых «трансформерами» (transformers).

Известные примеры моделей, использующих названную пару факторов (нейросетевая архитектура и увеличение обучающего корпуса и числа параметров нейросети), включают модель GPT-3 [Brown et al. 2020] — самую большую языковую модель на начало 2020 г., которая впервые показала, что увеличение количества обучаемых параметров до 175 млрд может обеспечить высокую точность решения многих заданий исключительно за счет «запоминания» примеров из корпуса, без какого-либо обучения конкретным навыкам («все уже сказано до нас»: на любой заданный вопрос в достаточно большом корпусе уже будет ответ).

Законы масштабирования таких моделей демонстрируют приблизительно линейное улучшение точности работы при увеличении числа параметров. Экстраполяция этой тенденции предполагает, что нейросеть с набором параметров больше на 3–5 порядков (секстиллион параметров) приведет нас к решению большинства поставленных сейчас задач обработки текста: любые типы классификации текстов, написание текстов в любом жанре и т. д. Эта наивная гипотеза важна для понимания верхней границы наших возможностей при исключительно экстенсивном подходе. Можно ли в рамках этого подхода действительно научиться решать любую задачу, сформулированную текстом? Есть ли фундаментальные интеллектуальные способности, которых недостает языковым моделям при таком подходе? Методы оценки, описанные далее, стремятся добиться ответа на эти и подобные вопросы.

2.1. Развитие методологии тестирования систем на основе теста Тьюринга

После появления теста Тьюринга [Turing 1950][10], представившего оценку способности машины к имитации человеческого интеллекта в виде переписки между машиной и судьями, возник широкий ряд смежных тестов интеллекта.

[8] Популярным золотым стандартом для оценки моделирования языка является корпус PennTreebank: рейтинг языковых моделей на нем доступен по адресу https://paperswithcode.com/sota/language-modelling-on-penn-treebank-word.

[9] https://catalog.ldc.upenn.edu/docs/LDC95T7/cl93.html.

[10] В ходе теста судьи взаимодействуют с собеседниками, каждый из которых может быть как компьютером, так и человеком. Задача судьи — понять, является ли собеседник человеком. Каждый участник стремится заставить судью признать, что он человек, на основании его ответов.

К подобным вариациям теста Тьюринга следует отнести в первую очередь

— схему Винограда (Winograd schema) — тест, содержащий текстовые вопросы о свойствах предметов и о привычных бытовых ситуациях, где правильный ответ обязательно требует снятия синтаксической неоднозначности [Winograd 1972];
— минимальный интеллектуальный Signal-тест (Minimum intelligent signal test, MIST) — вопросо-ответный тест, требующий от машины различных интеллектуальных навыков, знаний, логики, но при этом принимающий минимальную вариативность ответов — только «да» / «нет». Такой тест, предложенный в [McKinstry 1997], уменьшает субъективность судейства в оригинальном тесте Тьюринга, а также дает метрику «человечности» интеллекта системы — то есть доли правильных ответов;
— тест Тьюринга со специалистом (Subject-matter expert Turing test) — разновидность теста Тьюринга, при котором ответы машины должны задействовать экспертные специализированные знания, а правильные ответы не должны отличаться от ответов настоящих экспертов [McCorduck 2004];
— тест Эберта (Ebert test) — тестирование систем, включающих не только диалоговую письменную составляющую, но и синтез речи, при этом синтез речи должен быть достаточно хорош, чтобы судьи рассмеялись от шутки машины [Pasternack 2011].

Практика сравнения интеллектуальных способностей систем по результатам одного из таких тестов по-прежнему доминирует в современном исследовательском сообществе: системы могут сравниваться при игре в настольные и компьютерные игры, при решении прикладных задач. Однако широкая доступность больших данных для обучения и эффективность обобщений на таких данных у нейросетевых архитектур делают каждый из таких тестов уязвимым. Для объективной оценки интеллекта требуется больше, чем один результат системы на одном тесте. С одной стороны, удачное прохождение теста во многом зависит не от уровня сложности задачи, а от возможности предоставить системе неограниченный объем обучающих текстовых данных или предварительно заготовленной информации (корпусы научной литературы, справочники, энциклопедии): экспериментаторы могут не только вывести машину на произвольный уровень навыков, но и скрыть степень способности системы к интеллектуальному обобщению. С другой стороны, и искусственные, и естественные интеллектуальные системы дают нестабильные результаты при оценке на небольшом наборе заданий — повышение стабильности результатов требует увеличения объема и диверсификации тестов.

Подход, реализующий эту стратегию при оценке интеллектуальных систем, носит название бенчмаркинга (benchmark). Впервые он был представлен в работе [Fleming, Wallace 1986]: сравнение компьютерных систем в равных условиях требует аккуратной постановки задач и агрегации результатов. Бенчмарк-подход в применении к интеллектуальным системам подразумевает сочетание нескольких принципов:

1) фиксированное разделение данных: под сформулированную задачу собирается набор примеров, который фиксированным образом разделяется на три части — обучающую выборку, выборку для самопроверки участников и тестовую выборку для публичного сравнения систем (обычно в процентном соотношении 80 : 10 : 10 % или 70 : 15 : 15 % всех примеров);
2) закрытость тестовой выборки: «золотые» ответы на тестовые задания недоступны участникам и не могут быть угаданы путем перебора.

Текстовое представление интеллектуальных задач позволяет максимально разнообразно оценить способности соревнующихся систем, включая в задачи заведомую необходимость владения предметными знаниями (пчелы и самолеты летают за счет разных принципов аэродинамики), базовыми знаниями об объектах окружающей среды и их взаимодействии (зеленые сливы есть не стоит, желтые и красные уже созрели), логикой, способностью устанавливать причинно-следственные связи между описываемыми событиями.

Один из основных бенчмарков, идейно продолжающий формат вопросо-ответных вариаций теста Тьюринга, — это проект Stanford Question Answering Dataset, SQuAD [Rajpurkar et al. 2016]. SQuAD проверяет способность системы отвечать на связанные с пониманием прочитанного вопросы на английском языке: для каждого ответа на вопрос необходимо прочесть заданный абзац текста; система также может воздерживаться, когда задается вопрос, на который нет ответа в предоставленном фрагменте текста. Пример задачи[11]:

> Абзац: Warsaw is the capital and largest city of Poland. It stands on the Vistula River in east-central Poland, roughly 260 kilometres (160 mi) from the Baltic Sea and 300 kilometres (190 mi) from the Carpathian Mountains. Its population is estimated at 1.740 million residents within a greater metropolitan area of 2.666 million residents, which makes Warsaw the 9th most-populous capital city in the European Union. The city limits cover 516.9 square kilometres (199.6 sq mi), while the metropolitan area covers 6,100.43 square kilometres (2,355.39 sq mi).
>
> Вопрос: What is the largest city in Carpathia?
>
> Правильный ответ: <No answer>

SQuAD стал одним из первых бенчмарков, в котором в качестве точки отсчета использовался уровень решения задачи человеком (F-мера 89,45, при максимально возможном значении 100) и в котором эта точка отсчета была преодолена (в 2017 г. первые системы[12] преодолели порог точности извлечения информации из текста и превысили уровень человека в решении задачи). Актуальный рейтинг систем (пять лучших) на задаче SQuAD[13] представлен в таблице 1; всего на текущий момент выше средней оценки испытуемых поднялось 55 систем.

Таблица 1

Результаты лучших вопросо-ответных систем на задаче SQuAD

Позиция и дата	Модель	Метрики	
		EM	F1
1 (6 апр. 2020)	SA-Net on Albert (ensemble)	90.724	93.011
2 (5 мая 2020)	SA-Net-V2 (ensemble)	90.679	92.948
2 (5 апр. 2020)	Retro-Reader (ensemble)	90.578	92.978
2 (5 фев. 2021)	FPNet (ensemble)	90.600	92.899
3 (1 дек. 2020)	EntitySpanFocusV2 (ensemble)	90.521	92.824
	Средний результат человека	**86.831**	**89.452**

По приведенным в таблице данным видно, что результаты лидирующих систем в количественном отношении мало отличаются друг от друга, формируя плотный конкурентный рейтинг: такая ситуация встречается нередко в массовых соревнованиях по машинному обучению[14], однако для сложных задач с высокой «человеческой» планкой такой итоговый рейтинг стал неожиданностью.

[11] Источник: https://rajpurkar.github.io/SQuAD-explorer/explore/v2.0/dev/Amazon_rainforest.html.

[12] https://blogs.microsoft.com/ai/microsoft-creates-ai-can-read-document-answer-questions-well-person/

[13] На февраль 2021 г. Обновляемый рейтинг https://rajpurkar.github.io/SQuAD-explorer/.

[14] Примером могут служить соревнования на платформе https://www.kaggle.com/: задача классификации текста писем (спам / не спам) имеет более 800 решений, отличающихся сотыми

Системы, занимающие высокие места в рейтинге решения SQuAD, не известны нам до конца: про большинство из них нет научных публикаций. Однако из названий мы можем заключить, что часть из них основана на вариантах архитектуры encoder-decoder (ALBERT [Lan et al. 2019]), а также использует принцип голосования (ансамблирования — ensemble). Это значит, что обучению подвергаются несколько вариантов одной и той же нейросети; при тестировании каждая нейросеть пытается по предложению-вопросу предсказать правильное предложение-ответ. Затем из этих ответов выбирается самый частый, и он считается окончательным. Такой механистический подход позволил компьютерной системе отвечать на вопросы по текстам «Википедии» на 4 % лучше, чем человек.

К безусловным недостаткам теста SQuAD, послужившим причиной его быстрого преодоления, можно отнести, во-первых, однозначную формулировку задания. Здесь требуется дословный, с опорой на текст, ответ на явно сформулированный фактологический вопрос с вопросительным местоимением. Между тем типов вопросов гораздо больше, чем представлено в задаче («да / нет»-вопросы; вопросы, требующие выводов из текста и т. д.), но в систему оценки они не включены. Быстро справиться с этим тестом системе позволяет и очевидная простота имитации понимания при такой постановке задачи: чтобы правильно извлечь ответ из заранее предоставленного текста, необязательно учиться понимать текст, достаточно лишь научиться находить наиболее похожий или совпадающий с вопросом фрагмент в тексте и извлечь подходящую синтаксическую группу.

Следующий шаг в развитии и оценке интеллектуальных систем принадлежит бенчмарк-методологии, цель которой состоит в том, чтобы приблизить решение задачи понимания естественного языка. Эта методология получила название General Language Understanding Evaluation (GLUE) [Wang et al. 2018]. GLUE включает в себя усредненную оценку систем по многим задачам, требующим воспроизведения интеллектуальных способностей человека на основе текстовых данных, что делает их схожими со SQuAD, однако метод GLUE лишен обоих упомянутых недостатков SQuAD за счет более сложных и разнообразных заданий; собственно, как одно из 11 своих заданий он включает SQuAD.

Этот бенчмарк, изначально созданный на базе английского языка, уже был несколько раз воспроизведен на новом материале: усложненный вариант для английского [Wang A. et al. 2019], на китайском (CLUE) [Xu et al. 2020], русском (Russian SuperGLUE) [Shavrina et al. 2020b], польском (KLEJ) [Rybak et al. 2020], французском (FLUE) [Le et al. 2019] языках. Кроме того, он положил начало мультиязычному проекту XGLUE [Liang et al. 2020].

Методология GLUE включает следующие элементы:

1. Бенчмарк из девяти задач на понимание естественного языка, построенный на наборах данных, охватывающих весь диапазон данных разных размеров (от полных до небольших), жанров (от социальных медиа до художественной литературы) и включающих тексты различных степеней сложности. В наборах данных зафиксировано разбиение на обучающую, валидационную и тестовую выборку.
2. Набор диагностических данных, предназначенный исключительно для тестирования и анализа результатов обученных систем в отношении широкого спектра категорий, встречающихся на различных уровнях естественного языка (морфологическом, лексическом, синтаксическом, семантическом).
3. Общедоступный рейтинг систем для отслеживания уровня текущих решений и панель визуализации результатов.
4. Кодовая база (набор программ в открытом доступе) для быстрого воспроизведения результатов общедоступных систем.

Рейтинг решений GLUE агностичен относительно архитектур моделей, поэтому любая система, способная обрабатывать предложения и пары предложений и производить

и тысячными долями процента точности https://www.kaggle.com/c/spring20-cs189-hw1-spam/leaderboard.

соответствующие решения, имеет право участвовать в рейтинге. Рейтинг[15] первой версии представлен в таблице 2. Среднее значение по всем метрикам (в процентах), выстраивающее рейтинг, обозначено в третьей колонке («среднее»). Правее в таблице представлены метрики по следующим задачам:

1) определение корректных и некорректных предложений на базе The Corpus of Linguistic Acceptability (CoLa);
2) определение эмоциональной окраски предложений на базе The Stanford Sentiment Treebank (SST-2);
3) определение семантики предложений:
   а) являются ли два предложения парафразами или нет — на базе Microsoft Research Paraphrase Corpus (MRPC);
   б) являются ли предложения близкими по смыслу и насколько — на базе Semantic Textual Similarity Benchmark (STS-B);
   в) являются ли предложения парафразами — на базе вопросов Quora Question Pairs (QQP);
   г) соотнесение предпосылок и гипотез (есть ли между данной предпосылкой и гипотезой отношение следования, противоречия, нейтральности) — на базе MultiNLI (MNLI-M и MNLI-MM);
   д) схема Винограда (есть ли причинно-следственная связь между предпосылкой и гипотезой, для правильного ответа необходимо снять грамматическую неоднозначность) — на основе Winograd NLI (WNLI);
   е) лингвистическая диагностика (есть ли причинно-следственная связь между предпосылкой и гипотезой — пары минимально различающихся предложений с опорой на список из 33 признаков) (AX);
4) ответы на фактологические вопросы — на базе SQuAD (QNLI);
5) интерпретация семантики пар текстов: есть ли причинно-следственная связь между текстом с предпосылкой и гипотезой из текста — на базе Recognizing Textual Entailment (RTE).

Верхние пять позиций рейтинга GLUE (таблица 2, с 127) в настоящий момент занимают вариации архитектур на основе трансформеров (DeBerta, ENRIE и др.). Основной способ работы с ними включает два этапа: 1) обучение трансформера на большом корпусе (нейросеть учится предсказывать следующие предложения целиком или заполнять в них пропуски, штрафуется за ошибки, корректирует механизм внимания); 2) обученная на большом корпусе модель подвергается дообучению одному конкретному навыку: ей показывают обучающие примеры данных разных размеров, а научиться предсказывать нужно не следующий пример, а ответ в заданном формате (да / нет, метку класса, развернутый ответ). Этот второй шаг называют также дообучением или тюнингом (fine-tuning). Основная проблема тюнинга — возможность «забыть» вероятности, выученные на шаге 1: в таком случае модель не сможет должным образом применить знания, полученные на большом корпусе и обобщить навык, полученный на шаге 2. Однако, как видим в рейтинге, средний человеческий уровень при этом отошел на 15-е место, и трансформеры показывают результат выше на 3 %.

Подробный разбор заданий позволил получить первые оценки развивающегося «зоопарка» NLP-моделей[16] (model zoo). В бенчмарк были включены уже известные наборы заданий, в том числе низкоуровневые (классификация эмоциональной окраски, вопросо-ответные системы) и похожие между собой по формулировкам (задания на соотнесение предпосылки и гипотезы). Менее чем через год после выхода первой версии это привело к быстро возросшему уровню решений и сокращению в разбросе результатов среди конкурирующих систем: стало оправдано создание второй, усложненной версии — бенчмарка

[15] На февраль 2021 г. Обновляющийся рейтинг https://gluebenchmark.com/leaderboard.

[16] NLP — обработка естественного языка (natural language processing).

Таблица 2

Рейтинг систем GLUE

Место	Модель	Среднее	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	WNLI	AX	QNLI	RTE
1	DeBERTa Team — Microsoft	90.8	71.5	97.5	94.0/ 92.0	92.9/ 92.6	76.2/ 90.8	91.9	91.6	94.5	53.2	99.2	93.2
2	HFL iFLYTEK	90.7	74.8	97.0	94.5/ 92.6	92.8/ 92.6	74.7/ 90.6	91.3	91.1	94.5	52.6	97.8	92.0
3	Alibaba DAMO NLP	90.6	75.3	97.3	93.9/ 91.9	93.2/ 92.7	74.8/ 91.0	90.9	90.7	94.5	49.1	97.4	91.2
4	PING-AN Omni-Sinitic	90.6	73.5	97.2	94.0/ 92.0	93.0/ 92.4	76.1/ 91.0	91.6	91.3	94.5	51.2	97.5	91.7
5	ERNIE Team — Baidu	90.4	74.4	97.5	93.5/ 91.4	93.0/ 92.6	75.2/ 90.9	91.4	91.0	94.5	51.7	96.6	90.9
15	**Уровень человека**	**87.1**	**66.4**	**97.8**	**86.3/ 80.8**	**92.7/ 92.6**	**59.5/ 80.4**	**92.0**	**92.8**	**95.9**	**—**	**91.2**	**93.6**

Super General Language Understanding Evaluation (SuperGLUE). В таблице 3 приведены результаты англоязычных систем в рейтинге SuperGLUE, в таблице 4 — аналогичные результаты русскоязычных систем в рейтинге Russian SuperGLUE.

Все задачи SuperGLUE и Russian SuperGLUE разделяются на пять групп, сочетающих диагностику различных интеллектуальных способностей NLP-моделей:

1) здравый смысл: задачи PaRus, RUSSE, DaNetQA;
2) логический вывод: задач RWSD;
3) рассуждение: TeRRA, RCB;
4) машинное чтение: задачи RuCOs, MuSeRC;
5) общая лингвистическая диагностика: LiDiRus.

Подробное описание задач с русскоязычными примерами приведено далее.

Таблица 3

Актуальный рейтинг систем SuperGLUE

Место	Модель	Среднее	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
1	DeBERTa	90.3	90.4	95.7/ 97.6	98.4	88.2/ 63.7	94.5/ 94.1	93.2	77.5	95.9	66.7	93.3/ 93.8
2	T5 + Meena	90.2	91.3	95.8/ 97.6	97.4	88.3/ 63.0	94.2/ 93.5	92.7	77.9	95.9	66.5	88.8/ 89.9
3	**Уровень человека**	**89.8**	**89.0**	**95.8/ 98.9**	**100.0**	**81.8/ 51.9**	**91.7/ 91.3**	**93.6**	**80.0**	**100.0**	**76.6**	**99.3/ 99.7**
4	T5	89.3	91.2	93.9/ 96.8	94.8	88.1/ 63.3	94.1/ 93.4	92.5	76.9	93.8	65.6	92.7/ 91.9
5	NEZHA-Plus	86.7	87.8	94.4/ 96.0	93.6	84.6/ 55.1	90.1/ 89.6	89.1	74.6	93.2	58.0	87.1/ 74.4

Как можно заметить по таблице 3, в случае рейтинга SuperGLUE средний человеческий уровень решения задач преодолели две системы, с результатом выше на 0,5 %. При этом можно увидеть, что почти по всем отдельным задачам результат ИИ-систем остается ниже человеческого. Средний показатель лидирующих систем все же оказывается чуть-чуть

выше человеческого только за счет двух задач — ReCoRd и MultiRC; обе являются задачами на машинное чтение, где формат предписывает системе сделать выбор из вариантов ответа на вопрос по предоставленному тексту. Системы-лидеры T5 [Raffel et al. 2019] и DeBerta [He et al. 2020] обе основаны на архитектуре трансформеров: разработчики T5 представили концепцию обучения text-to-text, в котором они на стадии предварительного обучения модели поместили в обучающий корпус разнообразные наборы текстов и заданий с разметкой, ответами и соответствиями, как бы форсируя обучение модели конкретным навыкам заранее (teacher forcing). В случае с моделью DeBerta срабатывает новая методика тюнинга модели, при котором внутренние векторные представления текста намеренно портят специальным «шумом», чтобы сделать модель более устойчивой к формулировкам и заставить ее учиться на недостаточном объеме примеров.

Рейтинг моделей для русского языка (таблица 4) показывает, что уровень среднего испытуемого машинами пока не преодолен. Лучший результат (место 2) отстает на 13 % от уровня человека в среднем, хотя и оказывается выше него в отдельно взятых задачах MuSeRC и RuCoS. Эти задачи представляют собой полный аналог англоязычных задач машинного чтения MultiRC и ReCoRd, на которых трансформерные системы показывают наилучший результат и на английском языке. Архитектура Golden Transformer представляет собой инженерное решение: выбор ответа «большинством голосов» нескольких нейросетей encoder-decoder.

Таблица 4

Актуальный рейтинг систем Russian SuperGLUE

Место	Модель	Среднее	LiDiRus	RCB	PARus	MuSeRC	TERRa	RUSSE	RWSD	DaNetQA	RuCoS
1	**Уровень человека**	**0.811**	**0.626**	**0.68/ 0.702**	**0.982**	**0.806/ 0.42**	**0.92**	**0.805**	**0.84**	**0.915**	**0.93/ 0.89**
2	Golden Transformer	0.679	0.0	0.406/ 0.546	0.908	0.941/ 0.819	0.871	0.587	0.545	0.917	0.92/ 0.924
3	RuGPT3XL few-shot	0.535	0.096	0.302/ 0.418	0.676	0.74/ 0.546	0.573	0.565	0.649	0.59	0.67/ 0.665
4	MT5 Large	0.528	0.061	0.366/ 0.454	0.504	0.844/ 0.543	0.561	0.633	0.669	0.657	0.57/ 0.562
5	RuBERT plain	0.521	0.191	0.367/ 0.463	0.574	0.711/ 0.324	0.642	0.726	0.669	0.639	0.32/ 0.314

На рисунке показан пример решения текстовой задачи различными русскоязычными моделями: для правильного ответа необходимо произвести корректное разрешение анафоры.

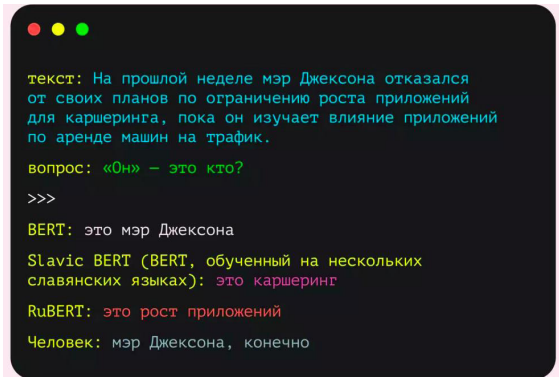

Рисунок. Решение задачи на логику (Russian Winograd Schema)

Далее приводится краткое описание каждой задачи.

2.1.1. Здравый смысл

PARus. «Выбор вероятных альтернатив» (Choice of Plausible Alternatives for Russian language, PARus) — это задача, которая направлена на оценку здравого смысла в причинно-следственных рассуждениях (commonsense causal reasoning). Каждый пример включает предложение-предпосылку с описанием ситуации и две альтернативные гипотезы — о причинах ситуации или о последствиях. Задача — выбрать наиболее правдоподобный из двух вариантов. Правильные варианты ответа перемешаны случайным образом: ожидаемая эффективность случайного угадывания составляет 50 %.

Пример:
Ситуация: Я получил скидку при покупке на кассе.
Тип вопроса: причина
Альтернатива 1: Я поздоровался с кассиром.
Альтернатива 2: У меня был скидочный купон.
Правильный ответ: альтернатива 2

RUSSE. Задача Word-in-Context (WiC) Russian SuperGLUE заимствует исходные корпусные примеры из проекта RUSSE[17] [Panchenko et al. 2018], материалы которого используются для обучения систем снятию неоднозначности у многозначных слов. Каждый пример состоит из пары предложений, в которых встречается одно и то же многозначное слово. Задача состоит в том, чтобы распознать, используется ли это слово в одном смысле или в разных (согласно результатам опроса носителей языка).

Пример:
Смысл указанного слова разный или одинаковый?
Слово: дорожка
Предложение 1: Бурые ковровые *дорожки* заглушали шаги.
Предложение 2: Приятели решили выпить на *дорожку* в местном баре.
Ответ: Смысл разный.

DaNetQA — это задача, которая, опираясь на методологию теста Minimum intelligent signal test и дизайн англоязычного аналога BoolQ, ставит перед системой «да/нет»-вопросы [Clark 2019]. Каждое задание в наборе состоит из одного абзаца текста и вопроса по его содержанию. Задача — дать на вопрос бинарный ответ (да / нет). Тексты для задачи взяты из «Википедии», а вопросы составлены с использованием краудсорсинга. Все вопросы были написаны авторами без каких-либо искусственных ограничений [Glushkova et al. 2020].

Пример:
Текст: В период с 1969 по 1972 год по программе «Аполлон» было выполнено 6 полетов с посадкой на Луне. Три экспедиции прошли без высадки астронавтов на Луне и шесть — с посадкой на Луне. Всего 24 астронавта США летали до Луны и обратно. Во время каждой из шести экспедиций с посадкой на Луне два астронавта выходили на поверхность Луны и один оставался в орбитальном модуле; таким образом, на Луне побывали 12 землян. Три астронавта — Джеймс Ловелл, Джон Янг и Юджин Сернан — по два раза летали к Луне, причем Янг и Сернан во время второго полета высаживались на Луне. Из-за аварии, произошедшей на «Аполлоне-13», Джеймсу Ловеллу не удалось высадиться на Луну.
Вопрос: Был ли человек на Луне?
Ответ: Да.

[17] https://russe.nlpub.org/2018/wsi/

2.1.2. Логический вывод

RWSD. Russian Winograd Schema — задача, посвященная снятию омонимии на основе логики и с опорой на базовые свойства объектов (форма, размер). Является переводным аналогом англоязычного набора заданий Winograd Schema [Levesque et al. 2012][18]. Схема Винограда состоит из предложений с синтаксической или морфологической неоднозначностью. Эта неоднозначность разрешается в предложениях при использовании знаний о мире. Корпус заданий составлен в соответствии с тестом Тьюринга. Одним из плюсов такой постановки задачи является простая форма машинного ответа (да / нет), а ответы систем делают даже для неспециалистов очевидным недостаток знаний у машин.

Пример:

Текст: Кубок не помещается в коричневый *чемодан*, потому что *он слишком большой*.
Есть ли связь: {антецедент: «чемодан», анафор: «он слишком большой»}
Правильный ответ: Нет! Это не чемодан слишком большой.

2.1.3. Рассуждение

TeRRA. Задача Textual Entailment Recognition for Russian (TERRA) посвящена обнаружению причинно-следственных связей между текстами, ответ предполагается в бинарной форме (есть связь или нет). Набор примеров был собран на основе правил и фильтрации корпуса Taiga [Shavrina, Shapovalova 2017] с последующей ручной редактурой. Задача охватывает основные потребности семантического вывода во многих приложениях NLP, таких как вопросо-ответные системы, информационный поиск, извлечение информации, реферирование текстов. Для решения задачи необходимо понять, следует ли значение одного текста из другого в данной паре.

Пример:

Предпосылка: Судно прибывает в сочинский порт за 7 часов. В то же время, билет на самолет до Сочи в летнее время стоит 3 тысячи, а время в пути — чуть меньше часа.
Гипотеза: Судном до Сочи добираться быстрее.
Правильный ответ: Нет, гипотеза не следует из предпосылки.

RCB. Russian Commitment Bank — это набор из пар текстов, в предложениях которых содержится или отсутствует причинно-следственная связь, а также размечены дополнительные дискурсивные характеристики: наличие вопроса, модальность, отрицание предшествующих условий. RCB относится к задачам распознавания причинно-следственных связей с делением на три класса (следование, противоречие, нейтральность). Так же, как и в задаче TERRa, каждый пример включает предпосылку и гипотезу. Однако в этом случае предпосылка представляет собой естественный фрагмент текста, а не одно предложение. Набор примеров был сформирован на материалах корпуса Taiga с помощью правил, написанных вручную, и проверен разметчиками.

Пример:

Предпосылка: «Мы не ожидаем, что это произойдет в ближайшие дни», — сказал чиновник.
Гипотеза: Это произойдет в ближайшие дни.
Правильный ответ: Нейтральность. Не ясно: может, произойдет, а может, нет.

[18] https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html

2.1.4. Машинное чтение

RuCoS. «Russian reading comprehension with Commonsense reasoning» (RuCoS) — это корпус заданий на понимание прочитанного, задания здесь составлены таким образом, чтобы обучить системы и протестировать их на наличие «здравого смысла» при чтении. RuCoS состоит из текстов новостных статей и их кратких содержаний, полученных автоматически; во фрагменте текста вместо одной из упомянутых сущностей ставится маркированный пропуск, который предлагается заполнить, выбрав из предложенных вариантов (сущностей, упомянутых в тексте). Целью RuCoS является оценка способности системы к выбору на основе базового понимания содержания документов, грамотного вывода из текста — что же произошло с упомянутыми в тексте сущностями. Задача разработана по методике англоязычного аналога ReCoRD [Zhang et al. 2018].

Пример:

Текст: Раз в пять лет в Давос съезжаются ученые со всего мира, чтобы произвести поверку приборов для измерения солнечной радиации. Метод архаичный, но действенный. Если бы только не облачность… Солнечная радиация — это неисчерпаемый источник энергии, залог жизни на Земле. В то же время она может представлять и весьма серьезную угрозу для обитателей нашей планеты — прежде всего, из-за влияния на климат. Поэтому точное измерение интенсивности солнечного излучения — и не вообще, а той его части, что достигает поверхности Земли, — имеет огромное значение — как теоретическое, так и практическое. Во Вселенной имеются самые разные типы звезд, и звезды эти находятся на разных стадиях эволюции. Изучение наиболее холодных белых карликов дает ученым возможность заглянуть в далекое прошлое нашей галактики.
На прошлой ассамблее в Праге Международный астрономический союз занимался преимущественно статусом Плутона и прочими проблемами классификации небесных тел. Теперь, в Рио-де-Жанейро, он повернулся лицом к науке.
Фрагмент с пропуском: «Поэтому в @placeholder прибыли и самые простые приборы, и высокотехнологичные изделия, предназначенные для установки на исследовательских космических аппаратах NASA», — говорит руководитель поверки Вольфганг Финстерле (Wolfgang Finsterle).
Варианты ответа: Давос / Земля / Плутон / Прага / Рио-де-Жанейро
Правильный ответ: Давос

Все текстовые примеры были собраны из открытых источников новостей, а затем автоматически отфильтрованы с помощью вопросо-ответных систем так, чтобы не допустить проникновения очевидных вопросов в набор данных (вопросы, на которые существующие системы не смогли дать ответ, были включены в набор). Затем тексты были отобраны по средней частотности[19] содержащихся слов и, наконец, проверены редактором вручную.

MuSeRC. Multisentence Reading Comprehension (MuSeRC) — задача, требующая от систем навыков совместной обработки предложений в тексте: задания представляют собой тексты и вопросы к ним, однако, чтобы ответить на вопрос, необходима информация из нескольких предложений. На вопрос невозможно ответить, не произведя логических операций на предложениях, не установив причинно-следственные связи между описываемыми событиями. К вопросам также прилагаются варианты ответов, каждый из которых не может быть найден в тексте непосредственно.

[19] Средняя частотность встречающихся в тексте слов должна составлять не менее 1 ipm (единицы на миллион) — условие, отбрасывающее тексты узкотематические, тексты с большим количеством терминологии, окказионализмов и т. п.

Пример:

Текст: (1) Экспансия коммерческой литературы сужает круг потенциальных читателей, которых в России осталось не так уж много. (2) Казалось бы, что за беда? (3) Читают — и пусть себе. (4) Всё лучше, чем пьянствовать. (5) Но не так-то всё просто. (6) Есть книги, без которых можно спокойно прожить. (7) Есть телевизор, есть газеты, есть компьютерные стрелялки. (8) А есть книги, без которых жить трудно. (9) И если в юности не попалась книга, перепахавшая душу, читатель для литературы потерян. (10) Он будет жевать литературный попкорн в полной уверенности, что читает книгу, не подозревая о том, что она всего лишь похожа на книгу, а к животворной литературе никакого отношения не имеет. (11) И таких читателей становится всё больше. (12) Но неужели всё так безнадёжно? (13) Неужели читателю, любящему живую книгу, остаётся утешаться нетленной классикой? (14) К счастью, нет. (15) Поразительная закономерность. (16) Живая книга чудом пробивается к читателю. (17) И диктат рынка ей не слишком большая помеха. (По В. Иванову).

Вопрос: Какая метафора используется в тексте?,

Варианты ответа:

1. Он будет жевать литературный попкорн в полной уверенности, что читает книгу, не подозревая о том, что она всего лишь похожа на книгу.
2. Тише едешь, дальше будешь, как закат.
3. Эта метафора — литературный попкорн.
4. Метафора корабля, несущего читателя к сокровищнице знаний по морю открытий.
5. Читая книгу, как сова в ночи, без пауз, учишься видеть и во тьме.

Правильные ответы: 1, 3

Набор данных состоит из 6 000 вопросов к 800 текстам из пяти разных источников: тексты для детей начальной школы, новости, художественные тексты, сказки, краткий пересказ содержания сериалов. Все тексты были собраны из открытых источников разных жанров и автоматически отфильтрованы по следующим параметрам: 1) длина абзаца, 2) количество именованных сущностей, 3) количество кореферентных связей. Вопросы к текстам задавали редакторы, опираясь на следующие принципы:

1. Ответ содержится в нескольких предложениях, а не в одном.
2. Ответ не прописан в тексте дословно.
3. Количество вариантов ответа может быть сколь угодно большим. Правильным / неправильным может быть любое количество ответов.

2.1.5. Общая лингвистическая диагностика: LiDiRus

Современные лингвистические подходы, описывающие различные поверхностные структуры в языках мира, используются для моделирования сложных заданий лишь частично из-за затратности получения корпусов с подобной лингвистической разметкой. Тем не менее небольшой набор данных, содержащий богатую разметку различных явлений синтаксиса, семантики, логики, используется для тестирования в методологии GLUE/SuperGLUE (представлен в обеих версиях)[20].

Linguistic Diagnostics for Russian (LiDiRus) — это диагностический набор данных для русского языка. Задания диагностики нацелены на проверку способности системы устанавливать причинно-следственную связь между парой предложений или решать,

[20] https://super.gluebenchmark.com/diagnostics

что связи нет. Пары предложений имеют специализированную разметку, которая охватывает широкий ряд лингвистических явлений, явно или неявно влияющих на качество распознавания причинно-следственной связи системой. Примеры диагностики составлены вручную, причем предложения внутри пары сформулированы таким образом, чтобы различия были минимальны: предложения различаются только одним или двумя лингвистическими свойствами (которые в одном предложении из пары изменены вручную).

Пример:
Предложение 1: Кошка сидела на коврике.
Предложение 2: Кошка не сидела на коврике.
Разметка: отрицание
Правильный ответ: Нет следования.

Разметка включает 33 лингвистических признака, от низкоуровневых (морфологических и синтаксических) до более высокоуровневых (включая уровни формальной семантики и знания об окружающем мире, знания свойств объектов и их взаимодействий).

Примеры всех категорий подробно описаны[21].

2.2. От GLUE к многоязычности

Успешное решения англоязычными системами бенчмарков GLUE и SuperGLUE положило начало распространению методологии тестирования на комплексную многоязычную оценку: проекты XTREME [Hu et al. 2020] и XGLUE [Liang et al. 2020] объединили имеющиеся материалы на разных языках для сравнения систем в задачах воспроизведения интеллектуальных способностей человека.

Бенчмарк XTREME представляет собой первый многоязычный проект для оценки ИИ-систем. XTREME охватывает 40 типологически разнообразных языков из 12 языковых семей и включает девять заданий, требующих анализа различных уровней синтаксиса или семантики. Включенные в XTREME задачи предполагают умение классифицировать тексты, анализировать морфологию и синтаксис, извлекать информацию и отвечать на вопросы на разных языках.

Согласно доступной информации, занимающая первую строку в рейтинге система VECO отстает от человека на 12 % и представляет собой encoder-decoder архитектуру, модифицированную под задачи перевода с языка на язык: модель учится по предложению на одном языке не предсказывать продолжение, а порождать это же предложение на другом языке. При обучении модели этой задаче — взвешивать взаимные соответствия между двумя последовательностями на входе (язык 1) и на выходе (язык 2) — в последовательности на языке 1 пропускается случайное слово, и модель должна научиться восстанавливать его с опорой на смысл на языке 2 [Luo et al. 2020]. Таким образом лучше выучивается межъязыковое соответствие; процедура повторяется с одной и той же моделью последовательно с разными парами языков. Система ERNIE (занимающая вторую строку в рейтинге) [Ouyang et al. 2020] основана на трансформерной языковой модели на базе похожего принципа: при обучении на большом корпусе модель учится предсказывать продолжение текста, но не целиком, а лишь заполняя искусственно поставленные в нем пропуски. Пропуски ставятся системой в случайном порядке — в разных местах предложения. Оба эти подхода являются экстраполяцией привычного нам механистического подхода на множество языков.

[21] https://russiansuperglue.com/ru/datasets/

Таблица 5

Уровень решения бенчмарка XTREME

Место	Модель	Среднее	Sentence-pair Classification	Structured Prediction	Question Answering	Sentence Retrieval
1	**Уровень человека**	**93.3**	**95.1**	**97.0**	**87.8**	—
2	VECO	81.1	88.6	75.4	72.4	92.1
3	ERNIE	80.9	87.9	75.6	72.3	91.9
4	T-ULRv2 + StableTune	80.7	88.8	75.4	72.9	89.3
5	Anonymous3	79.9	88.2	74.6	71.7	89.0
6	Polyglot	77.8	87.8	72.9	67.4	88.3

Стоит отметить некоторую неаккуратность, с которой авторы бенчмарка подошли к оценке уровня человека: аннотаторов-полиглотов задействовать не стали (разумеется, трудно найти людей, владеющих 40 языками) и взяли средний балл испытуемых из оригинальных задач, замеренный только на английском языке; среднюю оценку испытуемых в задаче определения части речи и вовсе взяли из классической работы про 3 % ошибок у разметчиков частей речи [Manning 2011], не проводя собственных замеров.

Бенчмарк XGLUE оценивает эффективность многоязычных предобученных систем в отношении их способности понимания и порождения текста на разных естественных языках. XGLUE состоит из 11 задач на 19 языках. Для каждой из них обучающие примеры доступны только на английском языке, и для успешного тестирования система должна обладать сильной обобщающей способностью с межъязыковым переносом знаний, чтобы использовать навыки, полученные на английских текстах, на других языках. По сравнению с параллельной работой проекта XTREME, XGLUE имеет два основных отличия. Во-первых, он включает межъязыковые задачи понимания и межъязыковые задачи порождения текстов одновременно — эти два типа задач формируют два отдельных рейтинга систем. Во-вторых, помимо пяти существующих межъязыковых задач (таких как поиск причинно-следственных связей), XGLUE также использует шесть новых задач из материалов поисковика Bing, включая классификацию новостей (NC), сопоставление запросов и объявлений (QADSM), ранжирование веб-страниц (WPR), сопоставление вопросов и ответов (QAM), генерацию вопросов (QG) и генерацию заголовков новостей (NTG). Такое разнообразие языков, задач и источников материала обеспечивает основу для оценки качества предобученных систем.

В XGLUE вовсе не приводится сведений о среднем балле человека в решении задач. Вопрос корректной оценки человеческого уровня в таком случае остается открытым, так как найти людей, способных решить задачи на равных условиях с компьютером, практически невозможно, и такой замер в любом случае не отражал бы «средний уровень» человека. Лучшая участвующая система, Filter [Fang et al. 2020], основана на языковой модели encoder-decoder с перестановками. Во время обучения распределению слов модели подаются как оригинальные «правильные» примеры, так и примеры с пометкой «неправильные», подвергшиеся случайным перестановкам слов; модель должна научиться отличать правильные от неправильных, получив отрицательный языковой материал. Система Unicoder [Huang et al. 2019], лидирующая в рейтинге генерации текста, представляет собой архитектуру encoder (без decoder), которая проходит по большому многоязычному корпусу с разметкой и учится решать задачу классификации и ответов на вопросы с фиксированным набором навыков.

Таблица 6

Уровень решения бенчмарка XGLUE

Задачи на понимание

Место	Модель	Среднее	NER	POS	NC	MLQA	XNLI	PAWS-X	QADSM	WPR	QAM
1	FILTER	80.1	82.6	81.6	83.5	76.2	83.9	93.8	71.4	74.7	73.4
2	Unicoder Baseline	76.1	79.7	79.6	83.5	66.0	75.3	90.1	68.4	73.9	68.9

Задачи на порождение текста

Место	Модель	Среднее	QG	NTG
1	Unicoder Baseline	10.7	10.6	10.7
2	MP-Tune	8.7	8.1	9.4

2.3. Открытые вопросы

Уровень систем, участвующих в описанных языковых тестах и соревнованиях, безусловно, вырос за последние десятки лет. Результаты систем, приведенные в предыдущей главе, приближаются к 100 %. Тем не менее, мы пока не наблюдаем появления сильного искусственного интеллекта. В связи с этим фактом перед сообществом стоит ряд открытых вопросов о балансе между экстенсивным и интенсивным развитием систем.

Технически запас экстенсивного увеличения нейросетевых систем подходит к границе доступного: для обучения трансформерных моделей в 2020 г. (как, например, в работе mT5 [Xue et al. 2020]) используется весь доступный материал самого большого интернет-корпуса / поискового индекса — Common Crawl[22]. Его объем текстов на 107 языках составляет 6,3 триллиона слов и знаков препинания (объем Brown corpus [Kučera, Francis 1967] — одного из первых электронных корпусов — 1,2 миллиона слов). Использующиеся в обучении систем корпуса, содержащие порядка десятков миллиардов и триллионов слов, имеют критически низкое качество материала (содержат дубликаты, спам и технический шум), не снабжены достоверными статистическими сведениями об источниках текстов, их жанровой атрибуции, информации о времени их написания и авторстве, а также часто не имеют стандартов описания и контроля версий (корпуса Oscar [Suárez et al. 2019], 166 языков, и C4 [Raffel et al. 2019], 101 язык).

Рост числа обучаемых параметров моделей также достиг триллионной отметки: система Switch Transformer [Fedus et al. 2021] представляет собой параллельно обучающийся ансамбль меньших моделей, совокупный объем которых превышает 1,6 триллиона параметров (первая нейронная языковая модель [Bengio et al. 2001] — миллионы параметров). Безусловно, стабильно удваивающиеся каждые два года вычислительные мощности процессоров обеспечивают исследовательское сообщество постоянно растущим объемом мощностей для работы с обучением таких больших языковых моделей (enormous language models), однако достигнутый масштаб делает их также и недоступными сообществу, а результаты — невоспроизводимыми.

Долгосрочная фиксация улучшений в работе систем моделирования языка сама по себе может быть названа проблемной из-за постоянной смены тестов.

[22] https://commoncrawl.org/

Однако с 1960-х гг. используется техническая метрика качества языкового моделирования — перплексия[23]. Перплексия (perplexity) — мера «удивления», определяющая, насколько хорошо модель предсказывает вариативность на тестовом корпусе. Чем меньше «удивление» на новых для модели текстах, тем лучше ее способность описывать вариативность. Для измерения перплексии используются фиксированные наборы текстов, в частности корпуса Penn Treebank [Marcus et al. 1994], WikiText-103 [Dauphin et al. 2017] и некоторые другие. На примере корпуса Penn Treebank заметен исторический прогресс в моделировании языка. Так, простая статистическая модель (smoothed 5-gram model, [Chen 1996]) достигает перплексии 141,2, тогда как рекуррентная нейросеть (LSTM, [Sutskever et al. 2014]) уже имеет меньшую перплексию, всего 82,7, а предобученная трансформерная модель Generative Pretrained Transformer-3 (GPT-3, [Brown et al. 2020]) «удивлена» тестом совсем мало — на результат 20,5.

Стоит отметить также, что общий рост ИИ-систем актуализирует проблему более строгой стандартизации обучающих корпусов. Поставленная корпусной лингвистикой проблема репрезентативности корпуса [Biber 1993] становится актуальной для оценки обучающих данных: вариативности и достаточности информации в корпусе, объема эксплицитно включенных энциклопедических материалов, необходимого объема диалогов, текстов различных жанров, источников, времени написания и т. д.

3. Заключение

Тестирование интеллектуальных систем на основании текстовых задач является распространенным способом в методологии оценки ИИ-систем и развивается с 1960-х гг., однако достояние лингвистики стало использоваться в формировании таких тестов сравнительно недавно — с появлением GLUE-методологии (2018 г.).

Бенчмарк-подход к оценке интеллектуальных систем на сегодняшний момент является доминирующим; он позволяет объединять оценки различных интеллектуальных способностей одной кумулятивной метрикой общего интеллекта. Интеллектуальные тесты, сформулированные в виде текстов, составляют основной метод оценки, позволяя формулировать самые разные типы заданий и сопоставлять уровень систем с человеческим интеллектом.

Методы бенчмаркинга активно применяются к множеству языков и сформировали направление оценки многоязычных систем. Системы сравниваются в том числе по способности переносить знания и навыки с одного языка на другой.

Текущие оценки ИИ-систем показывают, что для достижения человеческого уровня бывает достаточно увеличения обучающего корпуса и числа параметров нейросети. Составление новых бенчмарков вводит нас в междисциплинарную, пограничную зону между моделированием языка и моделированием интеллекта. Отделимость оценки одного от оценки другого возможна с формулированием новых типов текстовых бенчмарков, с контролируемым набором лингвистических свойств и условий решения.

За любым бенчмарком для машин пока еще стоит соревнование людей — лингвистов, программистов, математиков, философов. Ждем ли мы нового бума лингвистических задач — теперь для машин? Вполне вероятно, что именно «олимпиады» и «ЕГЭ» такого нового типа приведут нас к расцвету систем оценки ИИ, в том числе на русском материале.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

Baltrušaitis et al. 2018 — Baltrušaitis T., Ahuja C., Morency L. P. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, 41(2): 423–443.

[23] https://en.wikipedia.org/wiki/Perplexity

Bengio et al. 2001 — Bengio Y., Ducharme R., Vincent P. A neural probabilistic language model. *Advances in Neural Information Processing Systems 13 (NIPS 2000)*. Leen T. K., Dieterrich T. G., Tresp V. (eds.). Cambridge (MA): MIT Press, 2001, 893–899.

Biber 1993 — Biber D. Representativeness in corpus design. *Literary and Linguistic Computing*, 1993, 8(4): 243–257.

Brown et al. 2020 — Brown T. B., Mann B., Ryder N. et al. *Language models are few-shot learners*. Preprint, 2020. https://arxiv.org/abs/2005.14165.

Chen 1996 — Chen S. F. *Building probabilistic models for natural language*. Ph.D. diss., Harvard Univ., 1996. https://arxiv.org/abs/cmp-lg/9606014.

Chollet 2019 — Chollet F. *On the measure of intelligence*. Preprint, 2019. https://arxiv.org/abs/1911.01547.

Clark 2019 — Clark C., Lee K., Chang M. W., Kwiatkowski T., Collins M., Toutanova K. *BoolQ: Exploring the surprising difficulty of natural yes/no questions.* Preprint, 2019. https://arxiv.org/abs/1905.10044.

Dauphin et al. 2017 — Dauphin Y. N., Fan A., Auli M., Grangier D. Language modeling with gated convolutional networks. *Proc. of the 34th International Conf. on Machine Learning* (Sydney, 2017). Precup D., Teh Y. W. (eds.). = *Proceedings of Machine Learning Research*, 2017, vol. 70: 933–941.

Fang et al. 2020 — Fang Y., Wang S., Gan Z., Sun S., Liu J. *FILTER: An enhanced fusion method for cross-lingual language understanding*. Preprint, 2020. https://arxiv.org/abs/2009.05166.

Fedus et al. 2021 — Fedus W., Zoph B., Shazeer N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Preprint, 2021. https://arxiv.org/abs/2101.03961.

Fleming, Wallace 1986 — Fleming P. J., Wallace J. J. How not to lie with statistics: The correct way to summarize benchmark results. *Communications of the ACM*, 1986, 29(3): 218–221. https://doi.org/10.1145/5666.5673.

Kučera, Francis 1967 — Kučera H., Francis W. N. *Computational analysis of present-day American English*. Providence (RI): Brown Univ. Press, 1967.

Frankish, Ramsey (eds.) 2014 — Frankish K., Ramsey W. M. (eds.). *The Cambridge handbook of artificial intelligence*. Cambridge: Cambridge Univ. Press, 2014.

Glushkova et al. 2020 — Glushkova T., Machnev A., Fenogenova A., Shavrina T., Artemova E., Ignatov D. I. *DaNetQA: a yes/no Question Answering Dataset for the Russian Language*. Preprint, 2020. https://arxiv.org/abs/2010.02605.

He et al. 2020 — He P., Liu X., Gao J., Chen W. *DeBERTa: Decoding-enhanced BERT with disentangled attention.* Preprint, 2020. https://arxiv.org/abs/2006.03654.

Hu et al. 2020 — Hu J., Ruder S., Siddhant A., Neubig G., Firat O., Johnson M. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. *Proc. of the 37th International Conf. on Machine Learning (ICML)*. Daumé H. III, Singh A. (eds.). = *Proceedings of Machine Learning Research*, 2020, vol. 119: 4411–4421.

Huang et al. 2019 — Huang H., Liang Y., Duan N., Gong M., Shou L., Jiang D., Zhou M. *Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks.* Preprint, 2019. https://arxiv.org/abs/1909.00964.

Lan et al. 2019 — Lan Z., Chen M., Goodman S., Gimpel K., Sharma P., Soricut R. *ALBERT: A lite BERT for self-supervised learning of language representations*. Preprint, 2019. https://arxiv.org/abs/1909.11942.

Le et al. 2019 — Le H., Vial L., Frej J. et al. *FlauBERT: Unsupervised language model pre-training for French.* Preprint, 2019. https://arxiv.org/abs/1912.05372 2019.

Levesque et al. 2012 — Levesque H., Davis E., Morgenstern L. The Winograd Schema Challenge. *13th International Conf. on the Principles of Knowledge Representation and Reasoning*. Institute of Electrical and Electronics Engineers Inc. AAAI Press, 2012, 552–561.

Liang et al. 2020 — Liang Y., Duan N., Gong Y. et al. *XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation.* Preprint, 2020. https://arxiv.org/abs/2004.01401.

Luger, Stubblefield 2004— Luger G., Stubblefield W. *Artificial intelligence: Structures and strategies for complex problem solving.* 5th edn. San Francisco: Benjamin Cummings, 2004.

Luo et al. 2020 — Luo F., Wang W., Liu J., Liu Y., Bi B., Huang S., Huang F., Si L. *VECO: Variable encoder-decoder pre-training for cross-lingual understanding and generation.* Preprint, 2020. https://arxiv.org/abs/2010.16046v1.

Manning 2011 — Manning C. D. Part-of-speech tagging from 97 % to 100 %: Is it time for some linguistics? CICLing 2011: *International Conference on Intelligent Text Processing and Computational Linguistics*. Gelbukh A. F. (ed.). Dordrecht: Springer, 2011, 171–189.

Marcus et al. 1994 — Marcus M., Kim G., Marcinkiewicz M. A. et al. The Penn Treebank: Annotating predicate argument structure. *Human language technology*: Proc. of a Workshop held at Plainsboro, New Jersey, March 8–11, 1994. San Francisco: Morgan Kaufmann Publ., 1994, 114–119.

McCorduck 2004 — McCorduck P. *Machines who think*. 2nd edn. Natick (MA): A. K. Peters Ltd., 2004.

McKinstry 1997 — McKinstry C. Minimum Intelligent Signal Test: An alternative Turing Test", *Canadian Artificial Intelligence*, 1997, 41: pp. 35-47.

Nilsson 1998 — Nilsson N. *Artificial intelligence: A new synthesis*. San Francisco: Morgan Kaufmann Publishers, 1998.

Ouyang et al. 2020 — Ouyang X., Wang S., Pang C., Sun Y., Tian H., Wu H., Wang H. *ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora*. Preprint, 2020. https://arxiv.org/abs/2012.15674.

Panchenko et al. 2018 — Panchenko A., Loukachevitch N., Ustalov D., Paperno D., Meyer C., Konstantinova N. *RUSSE: The first workshop on Russian semantic similarity*. Preprint, 2018. https://arxiv.org/abs/1803.05820.

Pasternack 2011 — Pasternack A. (18 April 2011). "A MacBook May Have Given Roger Ebert His Voice But An iPod Saved His Life" [video]. Archived from the original on 6 September 2011. Retrieved 12 September 2011. https://www.vice.com/en/article/4xxa7j/a-macbook-gave-roger-ebert-his-voice-an-ipod-saved-his-life.

Poole et al. 1998 — Poole D., Mackworth A., Goebel R. *Computational intelligence: A logical approach.* New York: Oxford Univ. Press, 1998.

Raffel et al. 2019 — Raffel C., Shazeer N., Roberts A. et al. *Exploring the limits of transfer learning with a unified text-to-text transformer.* Preprint, 2019. https://arxiv.org/abs/1910.10683.

Rajpurkar et al. 2016 — Rajpurkar P., Zhang J., Lopyrev K., Liang P. *SQuAD: 100,000+ questions for machine comprehension of text.* Preprint, 2016. https://arxiv.org/abs/1606.05250.

Russell, Norvig 2003 — Russell S. J., Norvig P. *Artificial intelligence: A modern approach*. 2nd edn. Englewood Cliffs (NJ): Prentice-Hall, 2003.

Rybak et al. 2020 — Rybak P., Mroczkowski R., Tracz J., Gawlik I. *KLEJ: Comprehensive benchmark for Polish language understanding*. Preprint, 2020. https://arxiv.org/abs/2005.00630.

Searle 1980 — Searle J. Minds, brains, and programs. *Behavioral and Brain Sciences*, 1980, 3(3): 417–424. https://doi.org/10.1017/S0140525X00005756.

Shavrina, Shapovalova 2017 — Shavrina T., Shapovalova O. To the methodology of corpus construction for machine learning: "Taiga" syntax tree corpus and parser. *Proc. of "CORPORA-2017" International Conf.* Zakharov V., Belyaeva L. (eds.). St. Petersburg: St. Petersburg State Univ. Press, 2017, 78–84.

Shavrina et al. 2020a — Shavrina T., Emelyanov A., Fenogenova A. et al. Humans keep it one hundred: An overview of AI journey. *Proc. of the 12th Language Resources and Evaluation Conf.* (Marseille, 2020). Calzolari N. et al. (eds.). European Language Resources Association (ELRA), 2020, 2276–2284.

Shavrina et al. 2020b — Shavrina T., Fenogenova A., Emelyanov A. et al. *RussianSuperGLUE: A Russian language understanding evaluation benchmark*. Preprint, 2020. https://arxiv.org/abs/2010.15925.

Suárez et al. 2019 — Suárez P. J. O., Sagot B., Romary L. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache, 2019.

Sutskever et al. 2014 — Sutskever I., Vinyals O., Le Q. V. *Sequence to sequence learning with neural networks.* Preprint, 2014. https://arxiv.org/abs/1409.3215.

Turing 1950 — Turing A. Computing machinery and intelligence. *Mind*, 1950, vol. LIX, No. 236: 433–460.

Wang et al. 2018 — Wang A., Singh A., Michael J., Hill F., Levy O., Bowman S. R. *GLUE: A multi-task benchmark and analysis platform for natural language understanding.* Preprint, 2018. https://arxiv.org/abs/1804.07461.

Wang et al. 2019 — Wang A., Pruksachatkun Y., Nangia N., Singh A., Michael J., Hill F., Levy O., Bowman S. R. *SuperGLUE: A stickier benchmark for general-purpose language understanding systems*. Preprint, 2019. https://arxiv.org/abs/1905.00537.

Winograd 1972 — Winograd T. Understanding natural language. *Cognitive Psychology*, 1972, 3(1): 1–191. https://doi.org/10.1016/0010-0285(72)90002-3.

Xu et al. 2020 — Xu L., Hu H., Zhang X. et al. *CLUE: A Chinese language understanding evaluation benchmark.* Preprint, 2020. https://arxiv.org/abs/2004.05986.

Xue et al. 2020 — Xue L., Constant N., Roberts A., Kale M., Al-Rfou R., Siddhant A., Barua A., Raffel C. *mT5: A massively multilingual pre-trained text-to-text transformer*. Preprint, 2020. https://arxiv.org/abs/2010.11934.

Zhang et al. 2018 — Zhang S., Liu X., Liu J., Gao J., Duh K., Van Durme B. *ReCoRD: Bridging the gap between human and machine commonsense reading comprehension.* Preprint, 2018. https://arxiv.org/abs/1810.12885.

Получено / received 04.08.2021 *Принято / accepted 21.09.2021*