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

На правах рукописи

Шаврина Татьяна Олеговна

ЛИНГВИСТИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ И ОЦЕНКА ВЕКТОРНЫХ МОДЕЛЕЙ СЛОВ РУССКОГО ЯЗЫКА

Резюме
диссертации на соискание ученой степени
кандидата филологических наук

Научный руководитель:
кандидат филологических наук
Ляшевская Ольга Николаевна

Москва 2022

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

Публикации

На защиту выносятся перечисленные ниже пять статей, в двух из которых соискатель является единственным автором, в двух -- первым автором, и в одной - исполнителем в проекте.

1. **Шаврина Т.О**. О методах компьютерной лингвистики в оценке систем искусственного интеллекта. Вопросы языкознания. 2021. № 6. P.117-138. базы данных: Q2 Scopus

2. **T.Shavrina**. Word vector models as an object of linguistic research (Векторные модели как объект лингвистического исследования). Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2019”. 2019. P. 576-588. базы данных: Scopus

3. Alena Fenogenova, **Tatiana Shavrina**, Alexandr Kukushkin, Maria Tikhonova, Anton Emelyanov, Valentin Malykh, Vladislav Mikhailov, Denis Shevelev, Ekaterina Artemova. Russian SuperGLUE 1.1: Revising the Lessons not Learned by Russian NLP-models (Russian SuperGLUE 1.1: пересматривая невыученные уроки русскоязычных NLP-моделей). Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2021”. 2021. P .235-246. базы данных: Scopus

4. **Shavrina T.**, Fenogenova A., Emelyanov A., Shevelev D., Artemova E., Malykh V., Mikhailov V., Tikhonova M., Chertok A., Evlampiev A. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark (RussianSuperGLUE: бенчмарк оценки понимания русского языка), in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics. 2020. P. 4717-4726. базы данных: CORE A (Computer Science)

5. **Shavrina T.,** Emelyanov A., Fenogenova A., Fomin V., Mikhailov V., Evlampiev A., Malykh V., Larin V., Natekin A., Vatulin A., Romov P., Anastasiev D., Zinov N., Chertok A. Humans Keep It One Hundred: an Overview of AI Journey (На 100% человек: обзор решения AI Journey). Proceedings of The 12th Language Resources and Evaluation Conference Vol. 12. European Language Resources Association (ELRA). 2020. P. 2276-2284. базы данных: Scopus

Апробация работы

Основные результаты исследования были представлены на российских и зарубежных конференциях в период с 2019 по 2021 гг. в форме девяти устных докладов:

1. Конференция DIALOGUE 2021, 16 мая 2021, Москва, Россия
   Доклад: Russian SuperGLUE 1.1: Revising the Lessons not Learned by Russian NLP-models (Russian SuperGLUE 1.1: пересматривая невыученные уроки русскоязычных NLP-моделей)
2. Конференция DIALOGUE 2019, 29 мая 2019, Москва, Россия
   Доклад: Векторные модели как объект лингвистического исследования
3. Конференция Artificial General Intelligence, AGI-2020,
   13th International Conference, AGI 2020, 16–19 сентября 2020,
   Санкт-Петербург, Россия. Доклад: Russian SuperGLUE Creating a Language Understanding Evaluation Benchmark (Russian SuperGLUE: создание бенчмарка понимания русского языка)
4. The 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020, 16 – 20 Ноября 2020. Онлайн, Доминиканская республика.
   Доклад: RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark (RussianSuperGLUE: бенчмарк оценки понимания русского языка)
5. Симпозиум по анализу больших данных для выявления глобальных вызовов и трендов в сфере человеческого потенциала. Сессия 3 в рамках XXII Апрельской международной научной конференции НИУ ВШЭ. 12 апреля 2021, Москва, Россия. Институт статистических исследований и экономики знаний НИУ ВШЭ.
   Доклад: «Все способы измерить слона: Russian Superglue & RuSentEval»
6. XIII Шмелевские чтения (ПОВСЕДНЕВНАЯ РЕЧЬ КАК ОБЪЕКТ ЛЕКСИКОГРАФИИ), 23–25 февраля 2020 г., Москва, Россия
   Доклад: “Универсальные модели на корпусах повседневной речи как новые инструменты лексикографии” Институт русского языка им. В.В. Виноградова РАН
7. 4-й Колмогоровский семинар по компьютерной лингвистике и наукам о языке
   Доклад: Struggling with word vector models interpretation: some experience gained from basic linguistic practice (Проблемы с интерпретацией векторных моделей слов: опыт, полученный из лингвистической практики). Москва, Россия, Школа лингвистики ФГН НИУ ВШЭ.
8. Костомаровский форум, дискуссия «Компьютер говорящий и понимающий?»

Лаборатория когнитивных и лингвистических исследований, Институт Русского языка имени А.С.Пушкина, 25 мая 2021, Москва, Россия

9. Moscow HSE Pragmatics Workshop. 30 Сентября 2021, Москва, Россия
   Доклад: Russian Commitment Bank: machine learning lessons vs lessons of linguistics — all not learnt? (Russian Commitment Bank: уроки машинного обучения против уроков лингвистики — все не выучено?)

1. Тема, содержание и структура работы

Предлагаемая диссертация посвящена разработке методик лингвистической интерпретации и оценки векторных моделей слов для русского языка.

Векторные модели занимают существенное место в обработке естественного языка (Natural Language Processing, NLP), и являются неотъемлемой основой решения широкого круга задач, таких как классификация текстов (определение тематики текста, анализ эмоциональной окраски текста, классификация оскорбительных сообщений), извлечение информации (распознавание именованных сущностей, извлечение фактов), а также машинный перевод, суммаризации, а также генерации текстов (автоматическое создание уникальных текстов на заданную тему, в заданном стиле, жанре). Различные методы статистики и машинного обучения, включая нейросетевые, приводят к появлению таких артефактов, как векторные модели слов и текстов.

Векторные модели работают со словами и текстами в векторном пространстве признаков, ставя в соответствие тексту или слову численный вектор фиксированной длины. Как указывает (Conneau A. et al. 2018), одним из существенных затруднений в работе с векторными моделями является непрозрачность “черного ящика”, объясняемая способом обучения векторной модели. Признаки, представленные в векторах фиксированной длины, мало- и сложноинтерпретируемы, что существенно затрудняет сравнение и выбор лучшей векторной модели из имеющихся. Этот факт создает серьезную проблему для сравнения моделей, их результатов обучения и интерпретирумости, как формулирует работа (Rogers et al. 2020). Первые шаги в ее решении делаются для английского языка, прежде всего путем разработки методологии сравнительного анализа и бенчмаркинга (Wang A. et al. 2018, Wang A. et al. 2019). Благодаря развитию инженерной основы языкового моделирования в настоящий момент существуют сотни разновидностей различных векторных моделей слов и текстов, в том

числе большая часть из них адаптированы для применения на русскоязычном материалеl, например, модели проекта RusVectores (Kutuzov, Kuzmenko, 2017), а также DeepPavlov (Kuratov, Arkhipov, 2018).

Цель исследования

Создание методологии и инструментов своевременной оценки и интерпретации результатов обучения векторных моделей позволяет одновременно достичь двух важных целей, которым посвящено данное исследование.

- Во-первых, мы ставим целью разработать критерии для оценки векторных моделей моделирования, обоснованных с точки зрения теоретических представлений о языке. Эти критерии на годы вперед формирует повестку исследований, которые проводятся сообществом с целью улучшения текущих результатов моделирования языка. Разработка критериев оценки векторных моделей создает основу для развития как языковой инженерии, так и инструментов сравнения векторных моделей друг с другом и с носителями.
- Во-вторых, наша цель состоит в создании результатов обучения моделей, которые станут более понятными для человека, выявить дополнительные факторы влияния на качество моделирования языка. Инструменты интерпретации векторных моделей позволяют пролить свет на “черный ящик”, что делает востребованным разработку таких инструментов для русскоязычных моделей.

Для достижения приведенных целей в работе ставятся следующие **задачи**:

- проведение анализа существующих методов оценки векторных моделей для различных языков, поиск научных лакун в сфере теоретических требований к результатам языкового моделирования и практической оценкой результатов;
- исследования возможностей различных архитектур векторных моделей слов и текстов, включая актуальные модели на основе архитектуры трансформер, а также более старые модели дистрибутивной семантики (word2vec, GloVe), а также базовых моделей векторизации на основе коллокационных методов;

- создание набора новых тестов для проверки моделирования различных языковых интеллектуальных способностей, выраженных с помощью текстов: тестов на проведение причинно-следственных связей между событиями в текстах, тестах на логический вывод из текстов, тестов на общие и энциклопедические знания, тестов на снятие неоднозначности при помощи логики, а также тестов на машинное чтение — т. н. бенчмарка (benchmark) для русского языка;
- создание "лингвистической диагностики": набора диагностических тестов, определяющих влияние на результаты обучения различных явлений морфологии, синтаксиса, лексической и формальной семантики, а также непосредственно знаний о мире;
- подготовка кодовой базы, обеспечивающей инвариантность проведения тестов с моделью любой архитектуры (нейросетевой, дистрибутивно-семантической, правиловой, и т.д.);
- проведение тестирования существующих векторных моделей слов и текстов для русского языка в полученной системе оценки и интерпретации, анализ результатов, измерение среднего уровня человека в решении приведенных задач.

Актуальность исследования определяется двумя основными факторами:

- бурное развитие новых нейросетевых методов языкового моделирования предоставляет много новых объектов для изучения — моделей, которые необходимо сравнить между собой, выделить лучшие решения для дальнейшего прогресса в моделировании языка;
- отсутствие систем оценки и интерпретации векторных моделей слов и текстов для русского языка делает недоступным оценку и объективное сравнение моделей.

Вклад автора определяется следующими положениями: в работе (Shavrina T. 2019) автором единолично предложена методология интерпретации и сравнения статических векторных моделей, разработана кодовая база и методы тестирования обобщающей способности моделей в области лексической семантики. В работе (Shavrina T. et al. 2020b) представлено исследование, в котором автор руководил разработкой экспериментального ПО на основе векторных моделей, решающего варианты Единого Государственного Экзамена по русскому языку, включая тесты, задания с открытым

ответом и сочинение. Работа автора включает мотивацию и постановку задачи, методологию сбора решения и разработку решений 6 типов вопросов. В работах (Shavrina T. et al. 2020, Fenogenova A. et al. 2021) автору принадлежит разработка методологии для оценки и интерпретации векторных моделей, а также сбор первичных корпусных данных для последующей фильтрации и редактуры в подкорпусах задач. Работа (Шаврина Т.О. 2021) обобщает вышеназванные эксперименты и соединяет их в методологическом обзоре, описывающем методологические предпосылки, мотивацию принятых решений, а также текущие ограничения предложенной методологии.

Таким образом, в рамках данного исследования на защиту выносятся следующие положения:

1) постепенный прогресс векторных моделей слов и текстов измеряется с помощью набора разнообразных интеллектуальных задач, обеспечивая объективные зафиксированные равные условия для всех тестируемых моделей;
2) набор задач для тестирования языкового моделирования должен включать задачи, являющиеся достаточно сложными для актуального уровня развития прикладных языковых технологий; такой сложный уровень предлагает методология общего понимания естественного языка (General Language Understanding Evaluation, GLUE);
3) векторные модели демонстрируют способности к выявлению связей между постановкой интеллектуальных задач и явлениями языка, явно выражаемыми лексическими средствами. Например, *решать текстовые задачи на логический вывод с числом правильных ответов выше случайного, если в формулировке присутствует отрицание, дизъюнкция, конъюнкция или условная конструкция;*
4) выявления этих корреляций, однако, не достаточно, чтобы решать тесты без ошибок, ни одна из публично представленных векторных моделей для русского языка не приблизилась близко к уровню человека в решении представленных текстовых задач. С помощью представленного в работе инструментария фиксируются существенные ошибки и противоречия в моделировании языка, моделировании векторного пространства признаков слов и текстов у различных моделей.

Теоретическая значимость исследования определяется общим сближением достижений лингвистики и теории искусственного интеллекта, включая следующие факторы:

- в качестве основного инструмента для оценки уровня интеллектуальности систем представлены языковые тесты, затрагивающие морфологический, синтаксический, семантический, прагматический и дискурсивный уровни языка. Корпуса текстов в настоящий момент являются самым доступным способом обучения ИИ-систем и одновременно обладают высокой вариативностью, необходимой для формулировки самых разных интеллектуальных задач.
- впервые составлена и описана процедура тестирования интеллектуальных систем для русского языка, включающая обучение, валидацию и тестирование, а также подробный анализ результатов, диагностику ошибок и сравнение с уровнем человека.

Практическая значимость исследования и его влияние на дисциплину обусловлены представлением нового инструментария, бенчмарка Russian SuperGLUE (Shavrina T. et al. 2020a), состоящего из 9 новых корпусов интеллектуальных тестов для русского языка; каждый корпус тестов разделен на 3 фиксированные части — обучающую выборку, выборку для самопроверки участников, а также тестовую, с закрытыми золотыми ответами. Инструментарий доступен онлайн[1], под открытой лицензией. С момента публичного запуска онлайн-доступа к рейтингу (июнь 2020 года) процедуру тестирования и интерпретации прошли 1530 различных вариаций векторных моделей для русского языка; 23 из этих моделей представлены в публичном рейтинге[2] в сравнении с уровнем человека. В настоящий момент вышла вторая версия бенчмарка с дополнениями и коррекцией нескольких заданий первой версии (Fenogenova A. et al. 2021). 8 научных публикаций ссылаются на работу, согласно Google Scholar[3].

Новизна проделанной работы в рамках представляемого исследования представляется как совокупность теоретических и практических достижений в методологии бенчмарков на основе корпусов текстов.

2. Возможности векторных моделей слов и текстов

[1] https://russiansuperglue.com
[2] публичность результата в рейтинге определяется желанием автора системы. Рейтинг представлен по адресу https://russiansuperglue.com/leaderboard/2
[3] https://scholar.google.com/scholar?hl=ru&as_sdt=0%2C5&q=russian+superglue&btnG=

Векторные модели способны представлять слова и тексты в виде численных признаков, пригодных для обработки различными алгоритмами. Получаемые вектора признаков, соответствующие слову или тексту, могут использоваться для определения близких по смыслу слов, близких по тематике текстов, к ним также могут быть применены различным математическим операциям (Turney, Pantel, 2010): например, найти слово A, которое находится в таком же отношении к слову B, что слово C к слову D:

"Москва" — "Россия", "Сеул" —?

Ответ: "Корея"[4]

Векторные модели условно разделяются на две категории:

- статические, у которых вектор каждого слова или текста строго фиксирован и однозначно определен результатами обучения векторной модели на некотором корпусе текстов. К недостаткам таких моделей относят совпадение векторов признаков для омонимов и полисемичных слов, а также случайные вектора для самых частотных слов служебных частей речи, встречающихся в самых разнообразных контекстах;
- и динамические, или контекстуальные, при которых вектор признаков слова или текста зависит и может существенно меняться в зависимости от коллокатов слева и справа, являясь показателем контекстного значения.

К моделям первого типа (статические) относятся такие векторные модели, как

- простые коллокационные модели, vector space models на основе методов и корпусной статистики. Модели такого рода собирают частоты совместной встречаемости всех уникальных слов в корпусе: например, слово "лингвистика" встретилась в одном тексте со словом "компьютерная" 200 раз на 10 миллиардов слов, а "корпусная" встретилась в одном тексте со словом лингвистика 300 раз на 10 миллиардов слов. Так, для каждого слова собирается вектор длиной с размер словаря, где каждое число соответствует частоте встречаемости слова с каждым другим. Подобные вектора, безусловно, содержат множество нулевых элементов,

[4] На основании векторной модели word2vec, обученной на текстах НКРЯ и Википедии https://rusvectores.org/ru/calculator/#

а также имеют крайне большую размерность, так как число уникальных вхождений в словаре большого корпуса может равняться миллионам слов.

- нейронные модели дистрибутивной семантики: word2vec, fasttext, Glove и другие модели. Такие модели опираются на простые коллокационные модели, стремясь различными способами эффективно сжать вектора больших размерностей. В моделях дистрибутивной семантики чаще используются первичные частоты совместной встречаемости слов не в целом документе, а в небольшом контексте, например, на расстоянии 5 слов друг от друга. Эффективное сжатие больших векторов происходит за счет нейросетевых архитектур Continuous bag of words (CBoW) или Skip-gram (Mikolov et al. 2013). CBoW — архитектура, которая учится сжимать и восстанавливать вектор слова таким образом, чтобы предсказывает слово, исходя из окружающего его контекста. Skip-gram работает наоборот: по вектору текущего слова нейросеть учится предугадывать окружающие слова.

Второй тип моделей, динамический, в основном формируется за счет так называемых трансформерных моделей (transformers): моделей на основе архитектуры кодировщик-декодировщик (encoder-decoder) с механизмом внимания (attention) (Vaswani A. et al. 2017). Кодировщик нейросети принимает на вход текст, и механизм внимания взвешивает важность каждого слова, устанавливая коэффициенты важности — на основании них кодировщик формирует вектор контекста, а декодировщик решает заданную задачу — продолжает текст, или присваивает какую-то метку классификации. К таким архитектурам относятся, например, модели BERT (имеет только кодировщик), GPT-3 (только декодировщик), T5 (кодировщик и декодировщик) и другие.

3. Предлагаемая методика тестирования и интерпретации векторных моделей

Оценка статических векторных моделей

В работе (Shavrina T. 2019) статические векторные модели рассмотрены как самостоятельный объект лингвистического исследования. Подробно рассмотрены различные статические векторные модели русского и английского языка, их возможности и недостатки. Заключается, что с помощью статистических экспериментов над статическими векторами, полученными на различных корпусах русского языка, выделяются стабильные группы лексики с самыми однородными, стабильными

контекстами, независимо от жанрового и стилистического состава корпуса. Эти группы лексики включают прилагательные, обозначающие личные качества человека, национальность, профессии, топонимы, прилагательные времени.

В то же время наиболее нестабильной группой являются имена собственные — как наиболее редкие и контекстно-зависимые. Для русского языка был проведен эксперимент по оценке остаточного количества семантических и онтологических связей между известными парами слов, и качество моделей оценивалось на основе этого количества отношений, оставшихся в модели. Установлено, что слова из списка Сводеша более устойчивы к смене модели и сохраняют своих ближайших соседей гораздо чаще, чем слова из первой тысячи слов частотного словаря, а также чаще, чем случайные слова. Эти результаты также воспроизводятся и для английского языка.

В то же время, для анализа качества и интерпретации динамических векторных моделей нужна другая методология, пригодная для динамических векторов — она представлена в следующем разделе и подробно описана в (Shavrina T. et al. 2020a, Fenogenova A. et al. 2021).

Оценка динамических векторных моделей

Динамические векторные модели с момента их появления в 2016 году являются технологической основой большинства прикладных решений с самым высоким качеством. При помощи динамических векторных моделей впервые были по формальным метрикам были получены результаты выше среднего уровня асессоров: так, в задаче поиска ответа на вопрос в Википедии (задача Stanford Question Answering Datasets SQuAD, для английского языка (Li Yi, 2017)), на корпусе новостей превышено качество человеческого перевода с китайского на английский (Hassan H. et al., 2018)), также превышен уровень качества записи звучащей речи на слух текстом (английский).

Однако по этой причине стандартные прикладные задачи, такие как поиск ответов на вопросы в корпусе, классификация текстов по тематикам, эмоциональной окраске, извлечение именованных сущностей из текста и так далее, являются для объективного сравнения слишком простыми. При своей широкой представленности, прикладные задачи обработки русского языка не могут обеспечить существенный разброс метрик

между конкурирующими моделями, и часто решаются на уровне равном или выше уровня человеческого решения (95%+). В таком случае, между конкурирующими системами уменьшается разброс оценок, и их сравнение становится малоинформативным.

После появления теста Тьюринга (Turing 1950), представившего оценку способности машины к имитации человеческого интеллекта в виде переписки между машиной и судьями, возник широкий ряд смежных тестов интеллекта. Подробно эти методики рассмотрены в работе (Шаврина Т.О. 2021). Практика сравнения интеллектуальных способностей систем по результатам одного из таких тестов по-прежнему доминирует в современном исследовательском сообществе, однако, для повышения надежности результатов требуется диверсификации тестов.

Подход, реализующий эту стратегию при оценке интеллектуальных систем, носит название бенчмаркинга. Впервые он был представлен в работе (Fleming et al. 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 недоступны для внешнего перебора.Текстовое представление интеллектуальных задач позволяет максимально разнообразно оценить способности соревнующихся систем, включая в задачи заведомую необходимость владения предметными знаниями (пчелы летают не по тем же законам физики, что и самолет), базовыми знаниями об объектах окружающей среды и их взаимодействии (зеленые фрукты есть не стоит, желтые и красные уже созрели), логикой, способностью устанавливать причинно-следственные связи между описываемыми событиями.

Russian General Language Understanding Evaluation

Как статические, так и динамические векторные модели демонстрируют способность к решению относительно простых задач с заданными границами. Так, в работе (Shavrina T. et al. 2020b) показано, что с помощью векторных моделей возможно собрать 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, нейросетевую модель для генерации текста сочинения. 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 — в среднем 69 баллов из 100, что соответствует уровню "четверки".

Если в более простых задачах векторные модели демонстрируют свое превосходство, то с высокоинтеллектуальными задачами дело обстоит совсем иначе. Для более сложных интеллектуальных задач требуется хорошо разработанная методика определения степени текущих уровней решения задач.

Методика общего понимания естественного языка (General Language Understanding Evaluation, GLUE), впервые предложенная для английского языка, рассматривает оценку векторных моделей в комплексе: модели необходимо продемонстрировать уровень решения многих задач, желательно достаточно сложных, моделирующих различные интеллектуальные способности человека: предметные знания, логику, здравый смысл, способность к проведению причинно-следственных связей, демонстрацию понимания прочитанного текста. Эта методика оценивает пригодность модели к решению множества задач сразу, причем сами эти задачи наследуют методологию теста Тьюринга: включают в себя различные текстовые формулировки вопросов, обычно с вариантами ответов, и модели необходимо “притвориться человеком” - выбрать наиболее правильный вариант ответа.

Для русского языка эта методика интерпретативной оценки языковых моделей создается впервые и ложится в основу проекта Russian SuperGLUE. Проект содержит обновляемый рейтинг векторных моделей русского языка, их оценку на основе их ответов на вопросы, а также интерпретацию результатов на основе ошибок моделей, и корреляци ошибок с лингвистической информацией различных уровней - морфологии, синтаксиса, семантики, прагматики.

В рамках исследования впервые созданы корпуса интерпретирующих интеллектуальных задач для русского языка:

1. *Linguistic Diagnostic for Russian (LiDiRus):* оценка пригодности модели к проведению причинно-следственных связей на корпусе из минимальных пар предложений с искусственно усложненными формулировками и зафиксированными языковыми свойствами различных уровней.
2. *Russian Commitment Bank (RCB):* оценка пригодности модели к проведению причинно-следственных связей между событиями в новостных и художественных текстах;
3. *Choice of Plausible Alternatives for Russian language (PARus):* оценка пригодности модели к принятию решения на основе здравого смысла;
4. *Russian Multi-Sentence Reading Comprehension (MuSeRC):* оценка пригодности модели к причинно-следственным связям в прочитанном тексте;
5. *Textual Entailment Recognition for Russian (TERRa):* оценка пригодности модели к к проведению причинно-следственных связей в сравнении пар текстов;
6. *Russian Words in Context (RUSSE)*, оценка пригодности модели к снятию семантической неоднозначности на основе контекста и здравого смысла;
7. *The Russian Winograd Schema Challenge (RWSD):* оценка пригодности модели к решению логических задачи и целеполаганию;
8. *Yes/no Question Answering Dataset for the Russian (DaNetQA):* оценка пригодности модели к ответам на вопросы на предметное знания и понимание прочитанного текста.
9. *Russian Reading Comprehension with Commonsense Reasoning(RuCoS):* оценка пригодности модели к пониманию прочитанного текста.

Корпус (LiDiRus) был включен в общий список, так как имеет особенную задачу: лингвистическую интерпретацию. Лингвистическая интерпретация динамических

векторных моделей подразумевает исследование всевозможных зависимостей между выученными векторными признаками слов и текстов и известными лингвистическими параметрами, свойствами обучающих корпусов. Для этой цели LiDiRus формирует корреляционный анализ ошибок модели и различных явлений языка. Результатом процедуры является аналитический отчет по разнообразным ошибкам модели при наличии следующих свойств:

- Лексическая семантика: кванторы, именованные сущности, лексическое следование, симметрия, фактивность, морфологическое, отрицание, избыточность;
- Формальная семантика: отрицание и двойное отрицание, интервалы и числа, восходящая монотонность, нисходящая монотонность, немонотонность, различие в глагольном времени, конъюнкция и дизъюнкция, условные конструкции, универсальные и экзистенциальные предложения;
- Предикатно-аргументная структура: совпадение/несовпадение ролей ключевых аргументов глагола, предложные группы, наличие модификатора, способных определять не только сущность, к которой он относится, но и к другим, либо отменять определение такой сущности (интерсективность/ неинтерсективность), рестриктивность, анафора и кореферентность, согласование, активный/пассивный залог, эллипсис, номинализация, относительная клауза, дативные конструкции, генитив и партитив;
- Знания: здравый смысл, знания о мире.

Примеры всех свойств подробно описаны в Приложении 1.

Результаты для оценки и интерпретации русскоязычных моделей

Русскоязычная методология SuperGLUE подходит как для статических, так и для динамических векторных моделей слов.

К настоящему времени с помощью бенчмарка было оценено 1530 моделей, имеющих свои частные записи о производительности при выполнении различных интеллектуальных задач и подверженности моделей ошибкам, на которые влияют различные особенности языка. В таблице 1 вы можете увидеть лучшие результаты по производительности русскоязычной модели по сравнению со средней производительностью человека (к сентябрю 2021 года).

Таблица 1. Уровень человека и первые 3 векторные модели рейтинга, основанного на средней оценке на девяти интеллектуальных задачах. Общий балл рассчитывается путем усреднения результатов каждого задания. В результатах конкретной задачи используются следующие метрики: LidiRus - Matthews Correlation, RCB - F1 / Accuracy, PARus - Accuracy, MuSeRC - F1 / EM, TERRa - Accuracy, RUSSE - Accuracy, RWSD - Accuracy, DaNetQA - Accuracy, RuCoS - F1 / EM.

Название	Общий счет	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.ruRoberta-large finetune	**0.684**	0.343	0.357 / 0.518	0.722	0.861 / 0.63	0.801	0.748	0.669	0.82	0.87 / 0.867
3. Golden Transformer	**0.679**	0	0.406 / 0.546	0.908	0.941 / 0.819	0.871	0.587	0.545	0.917	0.92 / 0.924
4.ruT5-large-finetune	**0.634**	0.32	0.306 / 0.498	0.66	0.815 / 0.537	0.747	0.735	0.669	0.711	0.81 / 0.764

В настоящий момент ни одна из существующих моделей, прошедших тестирование, не приблизилась вплотную к уровню решения интеллектуальных задач человеком: общий счет в 81% правильных ответов достаточно сильно отстоит от лучшего результата моделирования (68% правильных ответов, модель ruRoBERTa large).

В рейтинге систем представлены 22 различных архитектуры, в том числе динамический векторные модели ruBERT, ruGPT-3, RuRoBERTa, их различные вариации и сочетания. Нижние строки рейтинга также представлены базовыми статическими решения:

- коллокационной моделью, полученной с помощью TF-IDF по корпусу Википедии,
- моделью случайного выбора ответа,
- моделью, отдающей всегда один тот же ответ.

Среди прочих, в рейтинге представлено правиловое решение, построенное на эвристиках: оно занимает 17ое место. В целом правиловые системы оказываются в нижней части рейтинга, в том числе решения, выбирающие всегда один и тот же ответ, а также решения, построенные на статических векторных моделях: так, решение, построенное на модели TF-IDF по Википедии, занимает 20-ую строчку и 43.4% правильных ответов из 100 возможных. В то же время такие динамические векторные модели, как ruT5, ruRoBERTa, ruBert занимают первые позиции в рейтинге после уровня человека (68.6%, 63.5%, 62% правильных ответов соответственно).

4. Заключение

В настоящем диссертационном исследовании предлагается новая методология оценки и интерпретации результатов обучения векторных моделей для русского языка.

Векторные модели слов показывают свою пригодность для решений в прикладных направлениях компьютерной лингвистики, таких как задачи классификации текстов, извлечение информации, машинный перевод. При рассмотрении их в качестве самостоятельного объекта исследования они требуют особых процедур для оценки и обеспечения корректного понимания предложений и текстов.

Обсуждаемая в исследовании методология предполагает, что для моделирования русского языка, приближенного к реальности, “естественный интеллект” асессоров и ИИ векторных моделей должны ставиться в равные условия тестирования. Методология включает в себя:

- набор из 9 новых корпусов с тестами на различные интеллектуальные способности, в том числе измеряющие отдельно качество предметных знаний, логику, причинно-следственные связи, понимание текста;
- набор лингвистической диагностики, проверяющий стабильность ответов и их правильность в зависимости от наличия в тесте различных явлений морфологии, синтаксиса, семантики;
- бенчмарки человеческого решения, решения популярными векторными моделями слов и текстов, а также эвристическими моделями.

Постепенное сближение методов компьютерной лингвистики и общего искусственного интеллекта считается взаимополезным: теоретическое представление об уровнях языка позволяет создавать системы тестирования и интерпретации векторных моделей, выделяя различные уровни усвоения языка моделями: морфологический, синтаксический, уровень лексической семантики, уровень формальной семантики, а также уровни владения базовыми понятиями языка.

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 методологии оценки общего понимания естественного языка (general language understanding evaluation). В рамках предлагаемой методологии качество моделирования языка оценивается через набор специализированных навыков, выражаемых через язык: владение причинно-следственными связями в тексте, логическим выводом и снятием неоднозначности, принятием решений в рамках описанных ситуаций, оперированием сведениями о базовых свойствах и признаках объектов повседневной жизни и абстрактных понятий.

Показывается, что предлагаемая методология представляет новый инструментарий для сравнения и интерпретации векторных моделей слов и текстов для русского языка. Прошедшие процедуру тестирования модели выстраиваются в общедоступный рейтинг с публичными результатами и отчетом о степени владения языком, о степени качества владения навыками и зависимости этого качества от различных явлений языка, заведомо представленных в примерах для тестирования. Составление рейтинга позволило выстроить целевые ориентиры для развития русскоязычных векторных моделей: за 2 года существования средний показатель качества на всех навыках у тестируемых моделей поднялся с 49.5% (у модели BERT-DeepPavlov, лучшее решение в момент запуска) до 75.5% (у лучшего текущего решения, модели Golden transformer). Безусловно, этот уровень владения навыками еще не достигает уровня носителей языка (81.1% в среднем); массовое тестирование показывает, что созданные модели также существенно отстают в задачах, где заведомо используется разнообразие явлений языка для усложнения формулировок задачи. Модели крайне чувствительны к изменениям формулировок задачи и использованию квантификаторов, числительных, конструкций с дательным падежом, эллипсиса, именованных сущностей — наличие этих неочевидных

факторов может существенно влиять на качество выучивания навыка моделями. Тем не менее, в задачах машинного чтения, в которых более надежный результат достижим при использовании большего объема данных при обучении, векторные модели показывают результат выше результата носителей языка.

Не вызывает сомнения, что методология может быть доработана, дополнена и обогащена новыми типами задач и разметки данных. Предложенные подходы показывают возможный вектор восполнения методологических лакун в лингвистической оценке языковых моделей и разработки более надежных, интерпретируемых практик работы с ними.

Составление потенциально новых интеллектуальных задач с использованием теоретических знаний о языке 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.

В настоящее время текстовые бенчмарки формируют мультидисциплинарную область, объединяющую такие направления, как лингвистика, машинное обучение, философия. Так называемое “лето искусственного интеллекта”, чье наступление связывается с развитием и популяризацией векторных моделей слов и текстов, требует от названных дисциплин новой методологии для фиксирования новых достижений, в том числе и систем на русском языке.

Список литературы

- Turney, P. D., P. Pantel. From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research, 37(1), 141-188. (2010)
- Mikolov, T., et al. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013).
- Pennington J., Socher R., Manning C. D. Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). – 2014. – С. 1532-1543.
- Conneau A. et al. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. arXiv preprint arXiv:1805.01070. – 2018.
- Kutuzov A., Kuzmenko E. (2017) WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In: Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science, vol 661. Springer, Cham
- Kuratov, Y., Arkhipov, M. (2019). Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. arXiv preprint arXiv:1905.07213.
- Wang A. et al. GLUE: A multi-task benchmark and analysis platform for natural language understanding //arXiv preprint arXiv:1804.07461. – 2018.

- Wang A. et al. Superglue: A stickier benchmark for general-purpose language understanding systems //arXiv preprint arXiv:1905.00537. – 2019.
- Vaswani A. et al. Attention is all you need //Advances in neural information processing systems. – 2017. – С. 5998-6008.
- Li Yi (2017) Avengers: Achieving Superhuman Performance for Question Answering on SQuAD 2.0 Using Multiple Data Augmentations, Randomized Mini-Batch Training and Architecture Ensembling. Stanford CS224N {Default} Project, Stanford University url
- Hassan H. et al. Achieving human parity on automatic Chinese to English news translation //arXiv preprint arXiv:1803.05567. – 2018.
- Xiong W. et al. Achieving human parity in conversational speech recognition //arXiv preprint arXiv:1610.05256. – 2016.
- Шаврина Т.О. О методах компьютерной лингвистики в оценке систем искусственного интеллекта. // Вопросы языкознания. 2021. № 6. С.117-138.
- Shavrina T. Word vector models as an object of linguistic research. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2019” Moscow, May 29–June 1, 2019
- Shavrina T., Fenogenova A., Emelyanov A., Shevelev D., Artemova E., Malykh V., Mikhailov V., Tikhonova M., Chertok A., Evlampiev A. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020. P. 4717-4726.
- Shavrina T., Emelyanov A., Fenogenova A., Fomin V., Mikhailov V., Evlampiev A., Malykh V., Larin V., Natekin A., Vatulin A., Romov P., Anastasiev D., Zinov N., Chertok A. Humans Keep It One Hundred: an Overview of AI Journey, in: Proceedings of The 12th Language Resources and Evaluation Conference Vol. 12. European Language Resources Association (ELRA), 2020. P. 2276-2284.
- Fenogenova A., ShavrinaT., Kukushkin A., Tikhonova M., Emelyanov A., Malykh V., Mikhailov V., Shevelev D., Artemova E.. Russian SuperGLUE 1.1: Revising the Lessons not Learned by Russian NLP-models (2021) A Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2021” Moscow, 2021.

Приложение 1

Лингвистическая диагностика Russian SuperGLUE.

Entailment

= Причинно-следственная связь, логический вывод, Natural Language Inference

В целом, мы рассматриваем проблему NLI на базе суждения о том, что среднестатистический читатель может посчитать явным после прочтения предпосылки, без учета возможной прагматики примеров. Неизбежно появление множества случаев, которые не являются чистыми, буквально очевидными, но мы хотим создать системы, которые будут способны делать те же выводы, что и люди. Это особенно справедливо в случае задачи целеполагания на основе "здравого смысла" (commonsense reasoning), который часто опирается на возможный

логический вывод. Мы стараемся исключить из данных сомнительные случаи и не используем никаких предложений, которые не являются грамматическими или семантически некорректными. В общем случае, используются стандарты, установленные в RTE Challenges, а также рекомендациями MultiNLI.

В каждой паре предложений (предпосылка и гипотеза), мы помечаем их одним из двух зависимых отношений:

- **Entailment**: гипотеза утверждает что-то, что определенно верно в отношении ситуации/события в предпосылке.
- **Not Entailment**: гипотеза утверждает что-то, что может быть правильным или определенно неверным в отношении ситуации/события в предпосылке.
  Эти определения по существу совпадают с теми, что были предоставлены разметчикам корпуса тестов MultiNLI. Создатели полагаются на предположение, что два предложения описывают одну и ту же ситуацию. Однако "ситуация" может включать несколько участников и действий, и степень детализации, при которой мы просим описанные ситуации быть одинаковыми, является несколько субъективной.

Russian Entailment:

- Написание кода на Java не слишком отличается от программирования в наручниках.
- Написание кода на Java подобно программированию в наручниках.

Russian - Not Entailment:

- Когда у вас есть снег, очень сложно обучиться зимним видам спорта, так что мы искали все способы изучить без снега то, что я мог бы потом повторить на снегу.
- Когда у вас нет снега, очень сложно обучиться зимним видам спорта, так что мы искали все способы изучить без снега то, что я мог бы потом повторить на снегу.

Лексическая семантика

Логический вывод

Причинно-следственное отношение может быть выражено не только на уровне предложения, но и на уровне лексики. Например, мы говорим, что "собака" лексически влечет за собой "животное", потому что все, что является собакой, также является животным, а "собака" лексически противоречит "кошке", потому что

невозможно быть одновременно обоими. Это относится ко всем частям речи (существительные, прилагательные, глаголы, многие предлоги и т. Д.), И взаимосвязь между лексическим и предложенным влечением была глубоко изучена, например, в системах естественной логики. Эта связь часто зависит от монотонности в языке, поэтому многие примеры Lexical Entailment также будут помечены одной из категорий Monotone, хотя мы не делаем этого в каждом случае.

- Falcon Heavy это самая маленькая ракета со времён ракеты-носителя NASA "Сатурн V", которая использовалась для полет на Луну в 1970-е годы.
- Falcon Heavy это самая большая ракета со времён ракеты-носителя NASA "Сатурн V", которая использовалась для полет на Луну в 1970-е годы.

Морфологическое отрицание

Это особый случай лексического противоречия, когда одно слово деривировано из другого: от доступного до недоступного, "пропорциональный" - "диспропорциональный" и т.д. Мы также включаем в эти примеры "когда" и "никогда". Мы также помечаем такие примеры как «Отрицание» или «Двойное отрицание», поскольку они могут рассматриваться как включающие логическое отрицание на уровне слов.

- Brexit — это обратимое решение, заявил сэр Майк Рейк, председатель WorldPay и экс-председатель группы BT, когда на прошлой неделе раздались призывы ко второму референдуму ЕС.
- Brexit — это необратимое решение, заявил сэр Майк Рейк, председатель WorldPay и экс-председатель группы BT, когда на прошлой неделе раздались призывы ко второму референдуму ЕС.

Фактивность

Пропозиции, появляющиеся в предложении, могут быть в любом причинно-следственном отношении в предложении, в зависимости от контекста, в котором они появляются.

- Все спичи – это политические публичные выступления.
- Джоан сомневается, что все спичи – это политические публичные выступления.

Во многих случаях это определяется лексическими триггерами (обычно глаголами или наречиями) в предложении. Например,

- Я признаю, что из X следует X
- Я не признаю, что их X следует X
- Я верю, что из X не следует X
- "Я отказываюсь делать X" противоречит "Я делаю X"
- "Я не отказываюсь делать X" не противоречит "Я делаю X"
- "Я почти закончил X" противоречит "Я закончил X"

- "Я только что закончил X" имеет следствие "Я закончил X" В русской лингвистической традиции такие конструкции определяются на основе импликативного типа глагола [А.Зализняк 1988]Конструкции, подобные примеру с "Я признаю", в датасете считаются фактивными, если предпосылка (пресуппозиция) сохраняется даже при отрицании. Конструкции, подобные вышеприведенным с "Я отказываюсь", в датасете считаются импликативными, и они чувствительны к отрицанию. Существуют также случаи, когда предложение (не) влечет за собой существование субъекта, упомянутого в нем, например,
- "Я нашел единорога" имеет следствие "Единороги существуют"
- "Я ищу единорога" не обязательно имеет следствие "Единороги существуют" Чтения, в которых сущность не обязательно существует, часто называют интенциональными прочтениями, поскольку они, похоже, имеют дело со свойствами, обозначенными описанием (его намерением), а не сводятся к набору сущностей, которые соответствуют описанию (его расширению, которое в случаях небытия будет пустым). Мы помещаем все примеры, связанные с этими явлениями, отметкой Factivity. Несмотря на то, что от контекста часто зависит, определяется ли вложенное суждение или существование сущности общим утверждением, очень часто оно сильно зависит от лексических триггеров, поэтому мы помещаем категорию в лексическую семантику.

Симметрия/Коллективность

Некоторые предложения обозначают симметричные отношения, а другие нет; например,

- "Джон женился на Мари" влечет за собой "Мари вышла замуж за Джона"
- "Джону нравится Мари" не влечет за собой "Мари нравится Джон"

Для примеров с симметричными отношениями действует правило, что их часто можно перефразировать, собирая оба аргумента в одну именную группу:

"Джон встретил Гари" влечет за собой "Джон и Гари встретились"
Является ли отношение симметричным, часто определяется по вершине группы - глаголу (например, "жениться" или "встречать"), поэтому мы классифицируем это явление в рамках лексической семантики.

- Республиканские законодатели просят президента Трампа воспользоваться противоречиями в структуре Белого дома в качестве основы для предстоящих дебатов Сената об иммиграционной политике.
- Президент Трамп попросит республиканских законодателей воспользоваться противоречиями в структуре Белого дома в качестве основы для предстоящих дебатов Сената об иммиграционной политике.

Избыточность

Если слово может быть удалено из предложения без изменения его значения, это означает, что значение слова было более или менее адекватно выражено предложением. Выявление таких случаев отражает понимание как лексической семантики, так и семантики всего предложения.

- Том и Адам громко шептались в театре.
- Том и Адам шептались в театре.

Именованные сущности

Слова часто называют сущности, которые существуют в мире. Есть много разных видов понимания, которые мы могли бы хотеть понять об этих именах, включая их композиционную структуру (например, "полиция Балтимора" - то же самое, что и "полиция города Балтимор") или их референты и аббревиатуры (например, "SNL" - это «Субботний вечер в прямом эфире». Эта категория тесно связана со знанием о мире, но фокусируется на семантике имен как лексических элементов, а не на базовых знаниях об их их денотатах.

- Стороны пришли к соглашению после их встречи в Европе.
- Стороны пришли к соглашению после их встречи в Стокгольме.

Квантификаторы

Количественное определение на естественном языке часто выражается через лексические триггеры, такие как каждый, большинство, некоторые и нет. Хотя мы резервируем категорию "Квантификация и Монотонность" для причинно-следственных связей, связанных с операциями над этими квантификаторами и их аргументами, мы решили рассматривать взаимозаменяемость квантификаторов как вопрос лексической семантики.

- Рассмотрим все слова из контекста как положительные примеры – и много негативных наугад из словаря.
- Рассмотрим некоторые слова из контекста как положительные примеры – и несколько негативных наугад из словаря.

Логика

Когда мы понимаем структуру предложения, часто появляется базовый набор небольших выводов, которые можно сделать с помощью логических операторов. Существует развитая традиция моделирования семантики естественного языка с использованием математических инструментов логики. Действительно, развитие математической логики первоначально было связано с вопросами о значении естественного языка, от аристотелевских силлогизмов до символов Фреге. Понятие причинно-следственной связи также заимствовано из математической логики. Поэтому неудивительно, что логика играет важную роль в языковом выводе.

Пропозициональная структура

Отрицание, Двойное отрицание, Конъюнкты, Дизъюнкты, Условные конструкции

Все основные операции логики высказываний появляются в естественном языке, и мы помечаем их там, где они соответствуют нашим примерам:

Отрицание: "Кот сидит на коврике" противоречит "Кот не сидит на коврике."

- Когда у вас есть снег, очень сложно обучиться зимним видам спорта, так что мы искали все способы изучить без снега то, что я мог бы потом повторить на снегу.
- Когда у вас нет снега, очень сложно обучиться зимним видам спорта, так что мы искали все способы изучить без снега то, что я мог бы потом повторить на снегу.

Двойное отрицание: На рынке невозможно не ориентироваться, влечет за собой "На рынке можно ориентироваться".

- Рынок станет сложнее, но на нём будет возможно ориентироваться.
- Рынок станет сложнее, но на нём не будет невозможно ориентироваться.

Конъюнкция: Температура и консистенция снега верны, значит и утверждение "температура верна" истино.

- Пациент несёт часть ответственности за успех лечения.
- Оба, и врач и пациент, несут часть ответственности за успех лечения.

Дизъюнкция: "Жизнь - это либо смелое приключение, либо ничто", включает "Жизнь - это смелое приключение", но не означает утверждение "Жизнь - ничто".

- Они слепо доверяют партнёрам.
- Реализуется одна из двух ожидаемых возможностей: либо они слепо доверяют партнёрам, либо у них конфликт интересов.

Условные конструкции: Условные конструкции немного сложнее, потому что их использование в языке не всегда отражает их значение в логике. Например, они могут использоваться на более высоком уровне, чем утверждение по вопросу: "Если вы думаете об этом, это идеальная тактика обратной психологии" влечет за собой "Это идеальная тактика обратной психологии"

- У Педро нет ослов.
- Если у Педро нет ослов, он не может их бить.

Квантификаторы

Универсальные, экзистенциальные Квантификаторы часто выражаются такими словами, как все, некоторые, многие и нет. Существует множество работ, моделирующих их значение в математической логике с помощью обобщенных

квантификаторов. В этих двух категориях мы фокусируемся на прямых выводах из естественных языковых аналогов универсальной и экзистенциальной квантификации:

Универсальные: Утверждение "У всех попугаев есть два крыла" влечет за собой, "Мой попугай имеет два крыла", но обратно не применимо.

- Ни у кого нет набора жизненных принципов.
- У каждого свой набор жизненных принципов.

Экзистенциальные: Утверждение "У некоторых попугаев есть два крыла", не значит, что "Мой попугай имеет два крыла".

- Никто не знает, как размножаются черепахи.
- Сьюзен знает, как размножаются черепахи.

Монотонность

Восходящая монотонность, Нисходящая монотонность, Немонотонность

Монотонность - это свойство позиций аргументов в определенных логических системах. В общем, это дает способ выведения причинно-следственных отношений между выражениями, которые отличаются только одним подвыражением. В языке это может объяснить, как некоторые выводы распространяются через логические операторы и квантификаторы. Например, обратите внимание, что "домашнее животное" влечет за собой "домашнюю белку", что влечет за собой "счастливую домашнюю белку". Мы можем продемонстрировать, как квантификаторы ∅, нет и ровно один отличаются по монотонности:

- "У меня есть домашняя белка" значит "у меня есть домашнее животное", но не значит "у меня есть счастливая домашняя белка".
- "У меня нет домашних белок" не влечет за собой "у меня нет домашних животных", но влечет за собой "у меня нет счастливых домашних белок".
- "У меня есть ровно одна домашняя белка" не значит ни "у меня есть ровно одно домашнее животное", ни "у меня ровно одна счастливая домашняя белка".

Во всех этих примерах домашняя белка появляется в том, что мы называем ограничителем положения квантификатора. Мы говорим: ∅ является восходяще монотонным в своем ограничителе: следствие в ограничителе приводит к следствию всего утверждения. нет - нисходяще монотонный в своем ограничителе: следствие в ограничителе приводит к обратному следствию всего утверждения. ровно один является немонотонным в своем ограничителе: следствие в ограничителе не приводят к следствию всего утверждения.

Таким образом, следствия между предложениями, которые строятся из следствий подфраз, почти всегда полагаются на суждения монотонности. Однако, поскольку это такой общий класс пар предложений, для сохранения значимости логической категории мы не всегда помечаем эти примеры монотонностью. Для аналогии - эти типы монотонности тесно связаны с ковариацией, контравариантностью и инвариантностью аргументов типов в языках программирования с подтипами.

Дополнительные логические структуры:

Интервалы/Числа, Время

Есть некоторые аспекты рассуждения более высокого уровня, которые традиционно моделировались с использованием логики; они включают в себя математические рассуждения, зависящие от чисел, и временные рассуждения.

Интервалы/Числа:"Я выпил больше двух бутылок сегодня вечером", подразумевает "Я выпил больше одной бутылки"

- Я не достиг своих целей в 1995 году.
- Я не могу достичь своих целей каждый год, начиная с 1997, а сейчас 2008.

Время: "Мэри ушла до того, как вошел Джон", влечет за собой "Джон вошел после того, как Мэри ушла".

- Джон вошёл после того, как Мэри ушла.
- Мэри ушла до того, как Джон вошёл.

Предикатно-аргументная структура

Важным компонентом понимания значения предложения является понимание того, как его части составлены в единое целое. В этой категории мы решаем проблемы по всему спектру, от синтаксической неоднозначности до семантических ролей и кореференции.

Синтаксическая неоднозначность: относительные клаузы, управление¶

Эти две категории имеют дело исключительно с разрешением синтаксической неоднозначности. Относительные клаузы и координационная сфера являются источниками значительной неопределенности в русском и английском языке.

- Мао был председателем Коммунистической партии с момента её прихода к власти в 1949 году и до своей смерти в 1976 году.
- Этот шаг знаменует конец системы, созданной Дэн Сяопином в 1980-х годах, чтобы предотвратить появление очередного Мао, который был председателем Коммунистической партии с момента её прихода к власти в 1949 году и до своей смерти в 1976 году.

Предложные группы

Связь предложной группы является особенно трудной проблемой, с которой синтаксические парсеры продолжают бороться. Мы рассматриваем это как проблему как синтаксиса, так и семантики, поскольку предложные группы могут выражать широкий спектр семантических ролей и часто семантически применяются за пределами их прямой синтаксической связи.

- В воскресенье у Джейн была вечеринка.
- У Джейн была вечеринка в воскресенье.

Основные аргументы

Глаголы управляют конкретными аргументами, особенно субъектом и объектом, которые могут быть взаимозаменяемыми в зависимости от контекста. Одним из примеров является эргативное чередование:

- "Джейк разбил вазу" значит "ваза разбилась"
- "Джейк разбил вазу" не значит "Джейк разбился".

Другие перестановки основных аргументов, такие как в Symmetry / Collectivity, также попадают под метку Core Arguments.

Чередования: активный/пассивный залог, генитив/партитив, номинализация, дативные конструкции

Все четыре из этих категорий соответствуют синтаксическим изменениям, которые, как известно, следуют определенным шаблонам на английском языке:

- активный/пассивный залог: "I saw him" (Я увидел его) равно "He was seen by me" (Он был увиден мной) и имеет следствие "He was seen" (Его увидели).
- генитив/партитив: "the elephant's foot" (нога слона) равно "the foot of the elephant" (слоновья нога).
- номинализация: "I caused him to submit his resignation" (Я вынудил его подать в отставку) влечет "I caused the submission of his resignation" (Я повлек его отставку).v
- дативные конструкции: "I baked him a cake" (Я испек ему торт) влечет "I baked a cake for him"(Я испек торт для него) и "I baked a cake" (Я испек торт), но не влечет "I baked him" (Я испек его).

Эллипсис

Часто аргумент глагола или другого предиката в тексте опускается (исключается), а читатель заполняет пробел сам. Мы можем построить примеры

причинно-следственной связи, явно заполнив пробел правильными или неправильными соответствиями. Например:

- Предпосылка: Путин настолько укоренился в правящей системе России, что многие из ее членов не могут представить себе другого лидера.
- Влечет за собой: Путин настолько укоренился в правящей системе России, что многие из ее членов не могут представить себе другого лидера, кроме Путина.
- Противоречия: Путин настолько укоренился в правящей системе России, что многие из ее членов не могут представить себе другого лидера, кроме себя. Это часто рассматривается как особый случай анафоры, но мы решили отделить эти случаи от явной анафоры, что часто также рассматривается как случай кореференции.

Анафора / Кореференция

Кореференция относится к тем случаям, когда несколько выражений ссылаются на одну и ту же сущность или событие. Он тесно связан с анафорой, где значение выражения зависит от другого (предшествующего) выражения в контексте. Эти явления имеют значительное совпадение, например, с местоимениями (она, мы, оно), которые являются анафОрами, которые ссылаются на свои антецеденты. Тем не менее, они также могут возникать независимо, например, корреляция между двумя определенными именными фразами (например, Тереза Мэй и премьер-министр Великобритании), которые относятся к одной и той же сущности. В эту категорию мы включаем только случаи, когда есть явная фраза (анафорическая или нет), которая совпадает с предшествующей или другой фразой. Мы создаем примеры для них почти так же, как и для категории эллипсиса.

- Джордж упал в воду.
- Джордж пошел к озеру, чтобы поймать рыбу, но упал в воду.

Интерсективность

Многие модификаторы, особенно прилагательные, допускают неинтерсективные употребления, которые влияют на их поведение в причинно-следственных связях. Например:

- Интерсекция: "он скрипач, и старый хирург" влечет за собой "он старый скрипач, и он хирург".
- Неинтерсективность: "он скрипач, и опытный хирург" не влечет за собой "он - умелый скрипач".
- Неинтерсективность: "он фальшивый хирург" не влечет за собой "он хирург"
  Как правило, интерсективное использование модификатора, как

"пожилой" у "пожилых людей", является тем, которое можно интерпретировать как обращение к набору сущностей с обоими свойствами (они пожилые и являются людьми). Лингвисты часто формализуют это, используя пересечение множеств, отсюда и берется название. Явление связано с фактивностью; например, подделка может рассматриваться как противоположный модификатор, и эти примеры будут помечены как таковые. Однако мы решили классифицировать интерсективность в структуре предикат-аргумент, а не в лексической семантике, потому что, как правило, одно и то же слово допускает как интерсективное, так и неинтерсективное употребление, поэтому интерсективность можно рассматривать как неоднозначность аргументной структуры.

Рестриктивность

Рестриктивность чаще всего используется для отсылки на свойство использования модификаторов существительных; в частности, рестриктивное использование модификатора - это такое использование, которое служит для идентификации описываемого объекта, тогда как нерестриктивное использование добавляет дополнительные детали к идентифицированному объекту. Различие часто может быть подчеркнуто последствиями:

- Рестриктивное: "я закончил всю домашнюю работу на сегодня" не означает "я закончил всю домашнюю работу"
- Нерестриктивное: "я избавился от всех этих надоедливых клопов" влечет за собой "я избавился от всех этих клопов".

Модификаторы, которые обычно используются нерестриктивно - это аппозитивы, относительные предложения, начинающиеся с того или иного (хотя они могут быть ограничительными, несмотря на то, что ваш учитель английского может вам сказать), и ругательства (например, надоедливые). Тем не менее, неограничительное использование может появляться во многих формах.

Неоднозначность в рестриктивности часто используется в определенных видах шуток.

Знание

Строго говоря, мировое знание и здравый смысл требуются на каждом уровне понимания языка для устранения неоднозначности смыслов слова, синтаксических структур, анафоры и многого другого. Таким образом, весь наш набор тестов в некоторой степени проверяет эти функции. Тем не менее, в этих категориях мы собираем примеры, в которых обоснование основывается не только на правильном устранении неоднозначности предложений, но и на применении дополнительных знаний, будь то конкретные знания о мировых делах или более обыденное знание о значениях слов, социум или физическая динамика.

Знание

В этой категории мы концентрируемся на знаниях, которые могут быть четко выражены в виде фактов, а также на более широких и менее распространенных географических, правовых, политических, технических или культурных знаниях. Примеры:

- "Это статья как будто из ИА Панорама" влечет "Эта статья читается как сатира".
- "Реакция была сильно экзотермической" влечет "Реакционная среда стала очень горячей".
- "Есть удивительные походы вокруг горы Фудзи" влечет "Есть удивительные походы в Японии", влечет "Есть удивительные походы в Непале".

Здравый смысл

В этой категории мы концентрируемся на знаниях, которые труднее выразить в виде фактов, и от которых мы ожидаем, что большинство людей будут обладать ими независимо от культурного или образовательного уровня. Это включает в себя базовое понимание физического и социального взаимодействия, а также лексического значения слов (помимо простого лексического вывода или логических отношений). Примеры:

- "Объявление об отъезде Тиллерсона вызвало шок во всем мире" противоречит "люди по всему миру были готовы к отъезду Тиллерсона."
- "Марк Симс встречается со своим парикмахером раз в неделю, в течение нескольких лет" влечет стрижку раз в неделю Марка Симса в течение нескольких лет.
- "Колибри действительно привлекают ярко-оранжевый и красный (следовательно, кормушки, как правило, такие цвета)" влечет "Желающие покормить колибри обычно окрашены так, чтобы привлечь их".

Приложение 2

Процедуры оценки модели

Submission ruGPT3 XL

Download submit | Edit | Make public | Delete

CORRECT

Jan. 27, 2021, 10:48 a.m.

Team: sberdevices

Total score: 0.534

Dataset	Score	Metric
LiDiRus	0.096	Matthew`s Corr
RCB	0.294 / 0.406	F1/Acc
PARus	0.676	Accuracy
MuSeRC	0.74 / 0.546	F1a/Em
TERRa	0.573	Accuracy
RUSSE	0.565	Accuracy
RWSD	0.649	Accuracy
DaNetQA	0.59	Accuracy
RuCoS	0.67 / 0.665	F1/EM

Diagnostic (Matthew`s Correlation): 0.096

Category	Score
LOGIC	-0.007250629987786855
KNOWLEDGE	0.059773386339694506
PREDICATE-ARGUMENT STRUCTURE	0.13997903121271693
LEXICAL SEMANTICS	0.18737835348183993

Diagnostic Details