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

На правах рукописи

Сериков Олег Алексеевич

ИНТЕРПРЕТАЦИЯ ПРЕДСТАВЛЕНИЯ ЗНАНИЙ В НЕЙРОСЕТЕВЫХ МОДЕЛЯХ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

Резюме

диссертации на соискание ученой степени
кандидата филологических наук

Научный руководитель:
кандидат филологических наук
Бонч-Осмоловская Анастасия Александровна

Москва — 2024

Введение

1.1 Тема, содержание и структура работы

Большие языковые модели являются флагманским подходом в решении задач автоматической обработки естественного языка и искусственного интеллекта. Такие модели снабжены механизмом отображения контекстов (последовательностей слов или токенов) в векторное пространство. Векторы контекстов оказываются информативными источниками признаков при решении разнообразных задач методами машинного обучения. Построенные таким образом системы показывают высокое качество на специфичных задачах[1] [1; 2] и сопоставимые с человеческими результаты в задачах общего искусственного интеллекта [3; 4]. В основе большого количества прикладных продуктов также лежат большие языковые модели [5; 6].

Одним из ключевых факторов, который делает большие языковые модели уникальными, оказывается их способность настраиваться, опираясь на большие коллекции данных. Благодаря большому количеству использованных в обучении данных (обучающие корпуса для языковых моделей могут достигать сотен гигабайт), языковые модели оказыаются способными корректно обрабатывать сложные лингвистические паттерны и порождать естественно звучащие тексты. Предварительное обучение на огромных корпусах текстов, оказывается достаточным для того, чтобы с минимальным дообучением моделей решать большое множество предварительных задач [7; 8]. Таким образом, предобученные языковые модели используются для решения широкого спектра задач, таких как анализ тональности, разметка именованных сущностей, классификация текстов и машинный перевод. Растёт популярность языковых моделей, предоставляемых в качестве отдельного сервиса (подобно системам Instruct GPT [9], GPT4 [10] и GigaChat [11]. Использование больших языковых моделей обеспечивает прирост качества и в алгоритмах обработки нетекстовых данных (видео, звук, табличные данные) для решения мультимодальных задач [12; 13].

[1] https://www.dialog-21.ru/evaluation/

Повсеместное применение языковых моделей вызывает особенный интерес к предсказуемости и обоснованию их поведения. Объяснимость поведения таких систем важна для оценки рисков их применения. Кроме того, способность систематизировать закономерности в поведении моделей потенциально полезна для их улучшения: обучение языковых моделей требует редких ресурсов, и, следовательно, опирается на разносторонний предварительный анализ. Используемые вновь и вновь системы обработки языка опираются в решении задач на когда-то усвоенное в форме миллиардов численных переменных знание. Такой формат представления информации затрудняет объяснение и предсказание их поведения. Таким образом, оказывается актуальной интерпретация представления знаний в нейросетевых моделях автоматической обработки естественного языка. Термин "знание" здесь и дальше мы понимаем в том же смысле, что он используется в методологии графов знаний, восходящей к реляционным семантическим структурам: пара <термы, набор семантических отношений между ними>, которая отражает экстралингвистическую реальность онтологического или энциклопедического характера. В экспериментах которые были реализованы в рамках диссертационного исследования знание будет представлено в качестве набора структур вида <тип отношения/предиката, его участники> или подобных им. Источником таких данных выбран ресурс Wikidata

Исследования, направленные на объяснение поведения построенных языковых моделей, ведутся в направлениях оценки и интерпретации поведения моделей на конкретных задачах, а также в направлении общей интерпретации таких систем при помощи техник пробинга (исследования связи между поведением отдельных элементов модели и каким-то свойством обрабатываемых данных). Большая часть пробинговых исследований сконцентрированы на анализе усвоения моделями грамматики языка, семантический пробинг осуществляется реже.

Целью данной работы является интерпретация лингвистических и экстралингвистических знаний, накапливаемых в языковых моделях.

Для достижения поставленной цели оказалось необходимо решить следующие **задачи**:

1. Анализ и систематизация релевантных направлений исследований, методов и инструментов интерпретации языковых моделей, их ключевых выводов и слабых сторон,

2. Определение требований к данным, необходимым для проведения интерпретационного исследования знаний в языковых моделях,
3. Составление методологии проведения интерпретационного исследования знаний в языковых моделях на основании проведенного обзора предметной области и установленных требований к данным,
4. Проверка применимости уровневой модели языка к анализу представления языка в нейросетевых моделях,
5. Проведение семантико-синтаксической интерпретации языковых моделей,
6. Проведение экстралингвистической семантической интерпретации языковых моделей,
7. Публикация инструментов интерпретации исследования знаний в языковых моделях в открытом доступе, реализующих воплощение методологии.

Основные положения, выносимые на защиту:

1. Среди существующих техник структурной интерпретации языковых моделей преобладающее большинство работ посвящены интерпретации грамматических структур, возникающих в моделях. Воплощение семантических структур и структур описания знаний в языковых моделях оказывается недоисследованным,
2. Для проведения исследований, восполняющих эту лакуну, оказывается полезна разметка скрытых в текстах структур. Источниками такой разметки могут быть как лингвистические ресурсы (например, скетчи с разметкой семантических ролей), так и ресурсы более общего порядка (графы знаний, соотнесённые с текстами),
3. Поведение различных участков нейронной сети оказывается скоррелированно с наличием различных типов предикатов, кодирующих знания. При этом для похожих по смыслу предикатов наибольшие корреляции показывают похожие участки нейронной сети,
4. Для обеспечения сопоставимости результатов интерпретации различных языковых моделей, относящихся часто к разным языкам, необходим инструментарий, стандартизующий проведение таких исследований.

Научная новизна:

1. Впервые для интерпретации языковых моделей применяются подходы, основанные на семантических ролях и графах знаний,
2. Впервые предложена методология эксплицитного анализа специализации голов внимания языковых моделей,
3. Иерархия типов предикатов, наполняющих граф знаний, была впервые установлена детерминированным путем на основании интерпретации языковой модели.

Научная значимость исследования обуславливается выработкой методологии применения богатого инструментария теоретической лингвистики к интерпретации семантической информации и знаний, выделяемых в языковых моделях. В научную значимость также вносит вклад выработка подхода к детерминированному подходу к построению иерархий типов отношений, присутствующих в графах знаний.

Практическая значимость исследования состоит в том, что были выложены в открытый доступ все инструменты, разработанные для проведения экспериментов. Таким образом для широкого сообщества исследователей становится возможным исследовать сильные и слабые стороны языковых моделей с точки зрения усвоения ими грамматического, семантического и экстралингвистического знания. Эксперименты, посвящённые интерпретации модели BLOOM, опубликованные вместе с описанием самой модели в работе [14] может служить примером такого использования разработанных инструментов. Также были выделены сильные и слабые стороны существующих мультиязычных языковых моделей, что может оказаться важным при выборе очередной модели для решения прикладной задачи.

Степень достоверности полученных результатов обеспечивается тем, что результаты работы были успешно опубликованы в изданиях и рецензируемых сборниках конференций. Результаты находятся в соответствии с результатами, полученными другими авторами, внося вклад в развивающуюся научную дискуссию об интерпретации нейросетевых моделей языка. Эксперименты, описанные в работе, опубликованы в статьях [15], [16], [14], [17], [18], [19].

Апробация работы обеспечена успешным применением полученных результатов научным сообществом в последовавших исследованиях [14].

1.2 Интерпретация языковых моделей

В последние годы исследования, направленные на интерпретацию языковых моделей, стали особенно популярны. В 2022 году на одной из крупнейших конференций, посвященных автоматической обработке естественного языка, EMNLP 2022 [20], именно направление интерпретации «черных ящиков» стало самым большим по количеству публикаций [21].

В работах [22] и [15] представлены обзоры области интерпретации языковых моделей с позиции отражающихся в них языковых структур. В обзорах рассматривается широкий спектр интерпретационных работ, посвященных грамматике языка и грамматике языков в типологическом разнообразии. Однако семантические ресурсы и графы знаний применяются в исследованиях гораздо реже. Этой проблеме посвящены исследования и разработки, описанные авторам в соавторстве в работах [15], [22], [17], [19], [15], [23], [16], [14], [18].

1.2.1 Обзор подходов к проблеме интерпретации языковых моделей

Многие из удивительных способностей больших языковых моделей состоят именно в том, что необходимые для решения практической задачи закономерности оказываются усвоены нейросетью до того, как она впервые используется для решения самой задачи [24]. В случае классического машинного обучения задачи решаются в другом порядке: сначала в данных размечаются зависимости, потом на них настраивается модель и лишь после этого она оказывается пригодной для решения целевой задачи. Одним из способов объяснить заблаговременное наличие полезных абстракций и системных характеристик в больших нейросетевых моделях языка является концепт самоорганизации. Этот концепт заключается в том, что сложные системы имеют тенденцию к вырабатыванию упорядоченных паттернов во взаимодействии их компонент [25]. Концепт самоорганизации используется для системного описания работы человеческого мозга [26], и поведения биологических видов, живущих в социумах, например, муравьев [27].

Значительная область текущих исследований посвящена интерпретации языковых моделей с точки зрения лингвистики. Мотивация таких исследований состоит в том, чтобы узнать, до какого порога модели “понимают” язык, или, точнее, до какого уровня обобщения, делаемые моделями о языке, совпадают с обобщениями и теоретическими построениями, принятыми в теоретической лингвистике. Важно отметить, что и концептуализация языка, производимая моделью, и концептуализации, принятые лингвистами являются лишь конструктами, и не имеют материальных аналогов в мире. Анализируя закономерности, определяющие поведение языковых моделей, исследователи часто прибегают к пробингу: поиску связи между поведением элементов нейронных сетей и свойствами обрабатываемых данных. Целью соответствующих исследований является установление уровня усвоения моделью конкретного уровня языка (например, морфологии, синтаксиса или дискурса в иерархической модели уровней языка [28]). Одним из возможных направлений таких исследований является анализа того, как именно лингвистические структуры представляют синтагматические и парадигматические механизмы языка в знании моделей. Иногда [29] исследователи предполагают, что способность к выделению таких структур говорит о приближении моделей к человеческому уровню мастерства на прикладных задачах.

Упорядочивая исследования о пробинговой структурной интерпретации языковых моделей, в [30] выделяется три этапа (большинство работ соответствуют первым двум этапам) в развитии пробингового исследования: “поведенческий”, “корреляционный” и “инвазивный” пробинг. Эти парадигмы естественным образом упорядочиваются. На этапе поведенческого пробинга исследователи устанавливают сам факт осведомленности языковой модели о каком-то свойстве. На этапе корреляционного пробинга в модели устанавливаются отдельные регионы — группы весов (чаще всего — целые слои весов) —, поведение которых заметно коррелирует с исследуемым свойством. Например, в работе [31] авторы выделили в модели компьютерного зрения нейроны, показывающие высокую активацию именно при обработке изображений с головами. На этапе инвазивного пробинга выделенные регионы подвергаются более пристальному исследованию: выделенные корреляционным пробингом веса модифицируются, и поведение полученной таким образом модели сравнивается с поведением исходной модели. Такой подход позволяет получить наглядное представление о

связи отдельных элементов языковой модели со свойствами обрабатываемых данных, однако в значительной степени оказывается комплементарным к корреляционному исследованию.

Пробинговые задачи позволяют измерить лингвистическую осведомленность регионов моделей, таких как слоев [32] или групп нейронов [33]. Измерение происходит следующим образом. При помощи внешней модели пытаются установить связь между предварительно размеченном набором данных, и записях о поведении на нём какой-то части нейросетевой модели. Чем легче такую связь установить, тем больше поведение исследуемой части нейросети коррелирует со свойством, размеченном в данных. Обычно внешней моделью выступает модель классификации, по поведению нейросетевой модели восстанавливающая аннотации в тексте. В таком случае внимание исследователя направлено на качество работы этой внешней модели. Внешняя модель называется пробинговой моделью. Процесс проведения подобных экспериментов называется пробингом. Наборы данных [34], сопровождающие языковые данные аннотациями для пробинга, называются пробинговыми наборами данных.

Существующие пробинговые наборы данных покрывают широкое разнообразие языковых признаков текстов, например, части речи токенов и деревья синтаксического разбора. Таким образом пробинговые эксперименты позволяют предположить, что искомая структура или искомое языковое свойство действительно заключено в определённых регионах модели. Это позволяет дополнительно охарактеризовать способность языковых моделей решать задачи, связанные с языком и именно лингвистикой [35—38]. Примерами исследований, опирающиеся на пробинговые методы, и указывающих на наличие лингвистических специализаций у участков нейронной сети являются работы [32; 34; 39—45]. В этих работах исследователи задались задачей установления того, как семантическое и грамматическое знание закодировано в языковых моделях. Синтаксический и морфологический пробинг включает в себя задачи, которые об установлении грамматической структуры текстов, по выработанным моделью векторным представлениям текстов. Семантические пробинговые задачи состоят в том, чтобы восстановить по векторному представлению словарное значение, семантическую связь или факт из базы данных фактов.

Ранние методы в этой области, опирающиеся на частеречные и морфологически аннотированные пробинговые наборы данных [39], позволили установить,

какая синтаксическая информация может быть закодирована в слоях нейросетевых моделей, способных обрабатывать язык. Появившиеся позже методы позволяют установить более явно, есть ли возможность восстановить целые деревья грамматического разбора текстов по внутренним представлениям предобученных моделей. В частности, в [44] представлен метод структурного пробинга для установления иерархических, подобных деревьям структур по векторным представлениям языка. Результаты применения метода к различным большим трансформерным моделям указывает на то, что некоторые промежуточные вычисления в таких моделях связаны с выделением иерархических структурных свойств (такая общая формулировка выбрана вследствие отсутствия явного соотнесения с теорией синтаксиса) текстов.

Отдельное направление работ посвящено пробинговым экспериментам с моделями на разных языках. В [46] были выделены синтаксические иерархические обобщения, сделанные моделями на разных языках при помощи метода структурного пробинга, а в работе [47] был предложен методологический фреймворк для мультиязычного морфосинтаксического пробинга, включающий в себя 15 различных задач, направленных на разные языки. Исследования на основе этого фреймворка позволили установить, что кросс-языковые типологические закономерности могут быть обнаружены при помощи пробинговых исследований

Пробинговая методология была также применена к определению семантической информации в векторных представлениях, выработанных отдельными компонентами моделей. Методы, основанные на восстановлении семантической аннотации по контексту, иллюстрируют способность контекстных представлений разделять несколько значений слова, представленного одним смысловым вектором [48]. При помощи подобного метода в [40] был поставлен послойный эксперимент о восстановления разметки семантических классов: более высокие слои моделей оказались лучше других в решении семантических задач. Пробинг связей, предложенный в [32], однако, показал, что семантические задачи приводят к более низким результатам в задачах пробинга, чем грамматические. Механизм внимания в нейросетях популярной сегодня архитектуры трансформера взвешивает все возможные связи между словами в предложении, вычисленные веса дальше используются для векторизации текстов и последующего решения задачи языкового моделирования. Разграничение влияния семантики

и синтаксиса на вычисление этих весов связей затруднено вследствие меньшего количества работ с семантическими пробинговыми задачами [23; 49]. Далее мы описываем подход к семантическому пробингу, который призван устранить такой дисбаланс.

1.2.2 Интерпретация представления знаний в языковых моделях

В этой работе мы прибегаем к семантическому пробингу, основанному на разметке семантических отношений. Семантические отношения и связанные ими термы описывают экстралингвистическую реальность. Такое энциклопедическое или онтологическое описание и формирует знание, исследуемое в нашей работе методами пробинга. Сформулированные на основе графов знаний семантические пробинговые задачи (задачи о семантических отношениях и их участниках) оказываются подобны синтаксическим (задачи о предикатах и их зависимых).

Прежде, чем осуществлять семантический пробинг, были проведены эксперименты [15] по обоснованию уместности уровневой модели языка в интерпретации моделей. Дальнейшее исследование соответствует поведенческому и корреляционному этапам трёхэтапной схемы пробинг-исследования. В работе [17] производится поведенческий пробинг с целью установления способности моделей к корректной обработке отношений, представленных в текстах. Сформированный на этом этапе набор данных оказывается, однако излишне гранулярным для проведения корреляционного пробинга. В работе [19] на основе графа знаний формируется более обширный набор данных, представляющих более простые семантические закономерности.

Инструменты, построенные для проведения экспериментов в работах [19], [17], [15] легли в основу фреймворка, позволяющего исследователям проводить разнообразные пробинговые исследования, затрагивающие отношения в нейронных сетях.

Обоснование уместности уровневой модели языка в пробинговом исследовании

В работе [23] был проведён пробинг в контексте уровневой модели языка. Корреляционные пробинг-задачи, представляющие морфологический, синтаксический и дискурсивный уровни языка были поставлены перед моделями на разных этапах их обучения. Такому хронологическому анализу подверглись две модели: BERT и T5, архитектуры кодировщик и кодировщик-декодировщик соответственно. 12 Пробинговых задач были взяты из существующих работ по пробингу SentEval [34], Morph Call [15], DisSent [50], DiscoEval [51], and BLiMP [52] для обеспечения сопоставимости результатов:

- Subject number (SentEval): грамматическая пробинг-задача бинарной классификации о числе подлежащего
- Лицо (Morph Call): грамматическая пробинг-задача бинарной классификации о наличии маркера лица
- Tree depth (SentEval): грамматическая пробинг-задача о глубине дерева разбора. Сформулирована в терминах классификации.
- Top constituents (SentEval): грамматическая пробинг-задача об определении главной составляющей в предложении
- Connectors (DisSent): дискурсивная пробинг-задача о восстановлении пропущенного между парами предложений предлога
- Sentence position (DiscoEval): дискурсивная пробинг-задача о правильном упорядочивании предложения среди четырёх других предложений
- Penn Discourse Treebank (DiscoEval): дискурсивная пробинг-задача о восстановлении дискурсивных отношений между сущностями Penn Treebank.
- Discourse coherence (DiscoEval): дискурсивная пробинг-задача об определении параграфов с перемешанными предложениями

Таким образом, анализ качества решения пробинговых задач позволяет сделать выводы в пользу уместности уровневой модели языка в пробинговом исследовании языковых моделей. Эксплицитное разделение морфологии и синтаксиса провести не удаётся, но последовательно разграничиваются грамматическое и дискурсивное знание моделей. Модели показывают подобные результа-

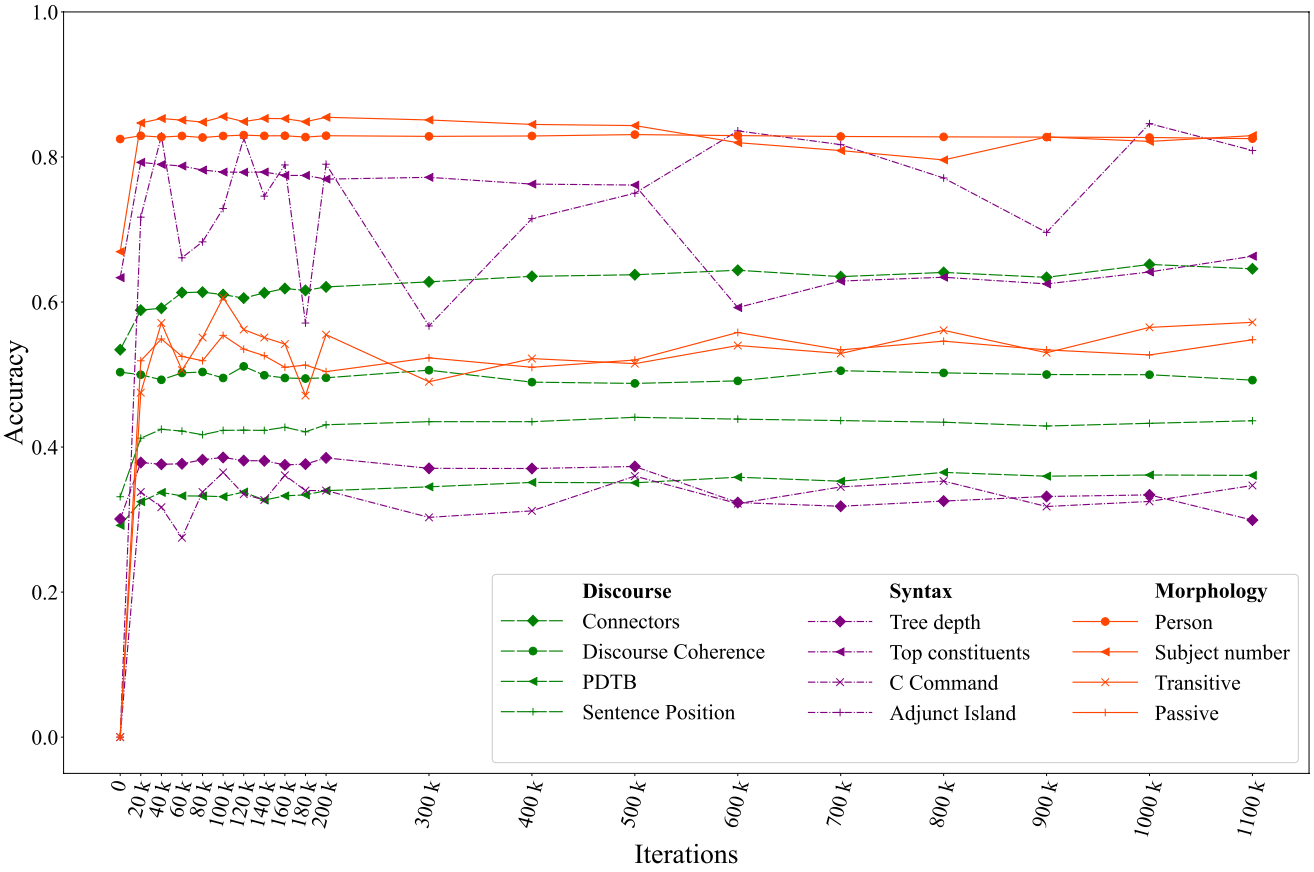

Рисунок 1.1 — результаты экспериментов об усвоении BERT уровней языка. Отображено качество решения моделями пробинговых задач на разных этапах обучения этих моделей. Грамматические задачи решаются хорошо начиная с примерно 100 тыс. итераций, а качество решения дискурсивных задач плавно улучшаетя на протяжении всего обучения моделей.

ты в решении задач несмотря на различия в архитектурах, и достигают итогового качества на грамматических пробинг-задачах в самом начале обучения: на протяжении первых 100000 итераций. В дальнейшем качество решения этих задач остаётся на достигнутом уровне. Задачи о дискурсе оказываются для моделей на порядок сложнее: не протяжении всего наблюдаемого периода обучения оно увеличивается плавно. Задачи Adjunct Island и Top constituents составлены неудачно, нестабильное поведение моделей на них может судить о неудачном выборе или количестве данных.

Грамматические уровни языка (синтаксис и морфология) оказались трудноразличимы при помощи пробинга языковых моделей, в то время как отделение грамматического и неграмматического уровней прослеживается. В дальнейшем мы исследуем как взаимодействие семантики с грамматикой, так и рассматриваем семантику в экстралингвистическом контексте.

Семантико-синтаксическая интерпретация языковых моделей

В рамках работы [17] было проведено соревнование, в котором участникам предлагалось автоматически, при помощи языковых моделей, отнести предложения к семантическим скетчам. Семантические скетчи (далее — скетчи) — это обобщенные лексикографические портреты, в которых зависимые слова классифицируются по грамматическим отношениям: для дочерних зависимых указывается, являются ли они определением, субъектом, объектом или соответствуют другой синтаксической роли. Представление такой информации в виде таблицы с популярными заполнителями синтаксических ролей позволяет составить семантический портрет соответствующего предиката.

Скетчи позволяют точнее исследовать вклад неядерных аргументов в предикат. Например, Локатив — типичный сирконстантный адьюнкт — является обязательным участником для глаголов с пространственным значением, таких как ‘находиться’. Именно локатив позволяет дифференцировать различные значения глагола быть, в то время как субъект оказывается гораздо менее информативным для проведения такого разделения. Семантические скетчи проиллюстрированы на Рисунке 1.2

State	Relation Correlative	Relation Relative	Object	Object Situation	Time
хороший было хорошо	человек был человек	это был это	человек был он	это было это	тогда было тогда
так было так	поэт быть поэтом	то есть то	лицо было лицо	оно был он	время был в то время
нужный было нужно	друг были друзьями	отец был он	отец был отец	дело было дело	год был в те годы
готовый был готов	то был то	человек был он	время было время	жить была жизнь	раньше было раньше
трудный было трудно	ребенок был ребенком	Но есть но	глаз были глаза	голос был голос	теперь был теперь
невозможность было невозможно	женщина быть женщиной	жить есть жизнь	женщина была она	разговаривать был разговор	жить был всю жизнь

Object Situation	Time	Object Situation Like	Locative	Modality	Time AlreadyStill
это было это	потом было потом	день был день	часы было на часах	на самом деле было на самом деле	уже было уже
что было что	вчера было вчера	вечер был вечер	исход были на исходе	нормальный будет нормально	еще было еще
дело было дело	давний было давно	ночь была ночь	двор был на дворе	наяву было наяву	все было все
оно был он	завтра будет завтра	утро было утро	тут было не тут-то	возможный было возможно	по прежнему было по-прежнему
то было то	год было в прошлом году	час было два часа ночи	гора была не за горами	невозможность было невозможно	все еще было все еще
случай был случай	поздний было позднее	весна была весна	СССР было в ссср	наверно было наверное	

Рисунок 1.2 — Семантические скетчи, соответствующие нелокативному и локативному значениям глагола быть

Проведение исследования об установлении способности моделей соотносить предложения со скетчами, соответствующим их вершинным предикатам, позволяет судить о наличии в моделях способности учитывать семантику в контексте связей между словами. Представленное участникам базовое решение сопоставляло предложения и скетчи при помощи поиска слов, зависящих от предиката предложения, в таблицах скетчей.

Это базовое решение, как и последовавшие решения, предложенные участниками, было основано на языковой модели RuBERT [53].

Алгоритм базового решения Для данного контекста *ctx*, скетч *sketch* выбирался на основании ранжирования, основанного на следующем механизме маскирования:

1. Сначала осуществлялся синтаксический анализ при помощи решения UDPipe ([54]), чтобы найти элементы, напрямую зависящие от вершинного предиката в предложении;
2. для каждого из зависимых, запоминались лучшие N кандидатов на замену Rep^N_{dep}
3. отобранные на прошлом этапе кандидаты пересекались по леммам $MLM^N_{ctx} = \bigcap\limits_{dep \in ctx} Rep^N_{dep}$
4. оценка $Score$ скетча вычислялась как количество токенов, встречающихся в пересечении скетчевого представления с запомненными заменами

$$Score\,(sketch, ctx) = \left| MLM^{1000}_{ctx} \cap Tokens_{sketch} \right|$$

Аналитика решений участников Системы, участвовавшие в соревновании, были основаны как на подходах, требовавших дообучения языковых моделей для решения задач, так и, что оказалось особенно релевантным с точки зрения интерпретации языковых моделей, системы, основанные на использовании систем без обучения вовсе.

Одна из систем участников игнорировала структуру скетчей, наивно превращая таблицу-скетч в последовательность, и векторизуя такой текст получала вектор скетча. Векторы скетчей сопоставлялись векторам предложений на основании косинусного расстояния между ними.

Другая система, превзошедшая первую, предложила подход, основанный на порождении предикатов по соответствовавшим им скетчам. Для генерации таких гипотез использовались шаблоны, мотивированные структурой семантических ролей в предикатах.

Для приведенных семантических скетчей сначала восстанавливался сам утаенный организаторами предикат. Для этого использовалась техника, заимствованная из классических методов разрешения омонимии — генерация по шаблонам. Примером такого шаблона может служить «[MASK] в школу», в качестве вероятного предиката на месте маски предлагающий глагол ходить. При помощи сопоставления полученных предикатов с предложениями восстанавливалось и само соотношение между предложениями и скетчами

Таким образом, оказывается, что модели действительно обращают внимание на семантические связи. Однако относительно слабый прирост качества

позволяет предположить, что семантические связи выделяются моделями с серьёзными отличиями от того, как они представлены в размеченном источнике данных.

Пробинг моделей для онтологий

Статья [19]представляет эксперименты и аналитику корреляционного пробингового исследования языковых моделей на примере модели BERT. Работа вдохновлена результатами [55], подхода к автоматическому отнесению слов к онтологиям. Авторы [55] иллюстрируют способность больших языковых моделей решать задачу разрешения омонимии, обращаясь к данным, восходящим к wikidata [56].

Для установления того, как реляционное семантическое и фактологическое знание располагается в языковой модели, был предложен подход, состоящий из нескольких этапов. На первом этапе был построен алгоритм, осуществляющий задачу извлечения отношений в текстах. Алгоритм нарочно основан на тривиальных подходах и состоит в последовательном решении двух задач классификации: сначала из всех возможных троек в тексте отбираются только потенциально представляющие семантически корректный триплет <субъект-предикат-объект>, затем отобранные тройки аттрибуцируются к непосредственно типам семантических отношений. В качестве векторизации троек токенов были испробованы как активации слоев модели, так и активации непосредственно “голов внимания” (attention heads), формирующих эти слои

Анализ компонент описанной системы позволил установить степень специализации каждой головы внимания на каждом типе отношений: более ранние слои модели оказались сильнее осведомлены о семантических отношениях. По поведению механизма межсловного внимания может удасться определить, находятся ли рассматриваемые токены в каком-либо семантическом отношении, и даже установить само это отношение. Головы внимания не имеют специализации на определенных типах отношений, оказываясь в той или иной степени информативными в задаче установления каждого из типов отношений. Тем не менее, именно это различие в степенях информативности голов внимания поз-

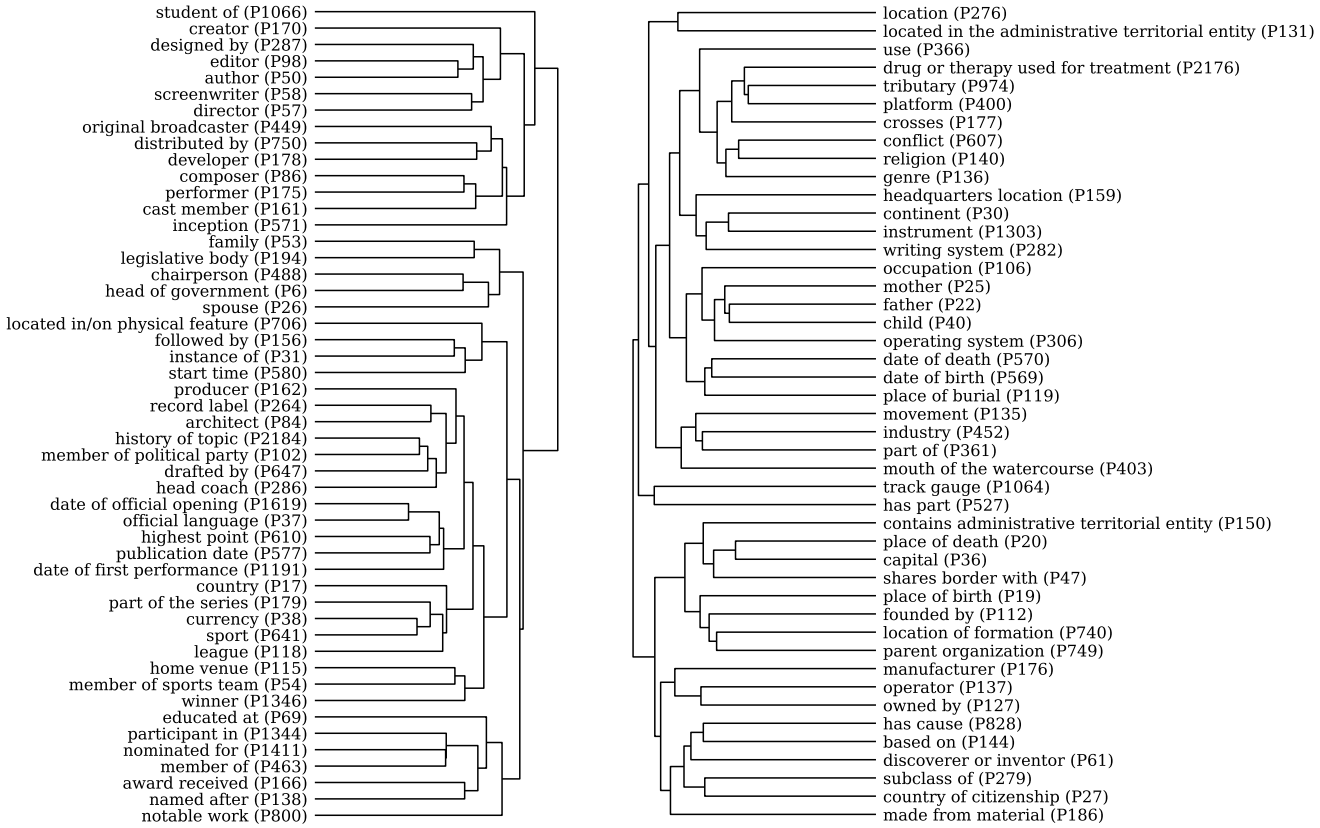

Рисунок 1.3 — Аггломеративная кластеризация отношений. В результате пробинга модели BERT в задаче определения семантических отношений были выделены области модели, особенно важные для определения каждого отношения. Сравнительный анализ важных для каждого отношения областей модели позволяет сгруппировать отношения по принципу похожести важных для них областей. Результирующее группирование представлено на рисунке в виде аггломеративной кластеризации отношений. Подобные по смыслу отношения оказываются определёнными в одну или близкие группы (пример: отношения founded by, location of formation, place of birth).

воляет сгруппировать типы отношений. Представление такой группировки в виде дерева аггломеративной кластеризации (Рисунок 1.2) указывает на тренд к объединению схожих типов отношений. Интересным сторонним наблюдением оказывается то, что иерархия отношений, представленная ресурсом Wikidata отличается от вычисленной описанным детерминированным образом. Тем не менее, использование информации о вычисленных нами иерархических связях для решения задачи выделения в тексте семантических отношений оказывается полезнее, чем предоставление информации об отношениях на основании Wikidata.

1.3 Заключение

В настоящем диссертационном исследовании представлена, применена и сделана доступной методология пробинга знания в нейросетевых языковых моделях. Сторонним эффектом разработки методологии стало также появление и вывод в широкий доступ инструментов, обеспечивающих возможность осуществлять пробинг языковых моделей

Пробинговое исследование знаний (семантических структур и отношений) проведено в два этапа. На этапе поведенческого пробинга выделена принципиальная важность для языковых моделей различных семантических ролей, и, так как модели контекстные, соответствующих им семантических отношений. Такое исследование проведено с привлечением набора семантических скетчей, сопровождающего дополняющего предикаты перечнями типичных заполнителей семантических ролей. Это позволило удостовериться в том, что если описать свойства текстов в терминах отношений и их участников, то между этими свойствами текстов и свойствами поведения моделей будет прослеживаться связь. На этапе корреляционного пробинга был произведена непосредственно интерпретация представления знаний в моделях. Вместо семантических скетчей именованные отношения брались из онтологической модели WikiData. Языковые модели были исследованы на предмет наличия в них структур, отвечающих за факты, изложенные в фреймворке двухместных семантических предикатов. Показано, что участков, имеющих избирательную специализацию нет как на уровне слоев языковой модели, так и на более подробном уровне голов внимания, составляющих эти слои. Анализ специализации голов внимания языковой модели на типах предикатов WikiData позволяет детерминированно выстраивать иерархию этих предикатов, и указывает на неконсистентность существующей субъективной иерархии.

Инструментарий, легший в основу методологии проведенного исследования, оказался удобен для проведения других исследований по интерпретации поведения языковых моделей и был представлен в виде системы для интерпретации языковых моделей, изложенной в опубликованной статье [16]. Система позволяет исследователям проводить большие количества экспериментов (сотни и тысячи) об интерпретации языковых моделей и анализировать их результа-

ты, представленные в едином формате и интерфейсе. Система была применена при интерпретации новой языковой модели BLOOM (эксперименты изложены в опубликованной статье [14]) и позволили установить пределы обобщения грамматических способностей языковой модели на незнакомые ей языки.

Существующие большие языковые модели становятся основой все большего количества прикладных систем, так и исследований. Практическое применение этих моделей и их исследование оказывается затруднено невозможностью объяснения точных причин того или иного результата работы таких моделей. Одной из техник упорядочивания такой степени неопределенности является пробинговая интерпретация в контексте известных структурных описаний области применения моделей. Структурные описания знаний, однако, прежде не использовались в контексте интерпретации моделей. Предложенная и аппробированная в работе методология пробинговых исследований предлагает теоретическую базу и инструментарий для будущих исследований.

Список литературы

1. RUSSE-2022: Findings of the First Russian Detoxification Shared Task Based on Parallel Corpora / D. Dementieva [и др.]. —.
2. RuArg-2022: Argument Mining Evaluation / E. Kotelnikov [и др.] // arXiv preprint arXiv:2206.09249. — 2022.
3. Superglue: Learning feature matching with graph neural networks / P.-E. Sarlin [и др.] // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2020. — с. 4938—4947.
4. RussianSuperGLUE: A Russian language understanding evaluation benchmark / T. Shavrina [и др.] // arXiv preprint arXiv:2010.15925. — 2020.
5. *Блог Яндекса*. Алгоритм «Палех»: как нейронные сети помогают поиску Яндекса. — доступ 3 ноября 2023. — https://yandex.ru/blog/company/algoritm-palekh-kak-neyronnye-seti-pomogayut-poisku-yandeksa.
6. Conversational ai: The science behind the alexa prize / A. Ram [и др.] // arXiv preprint arXiv:1801.03604. — 2018.
7. Language models are unsupervised multitask learners / A. Radford [и др.] // OpenAI blog. — 2019. — т. 1, № 8. — с. 9.
8. Language Models are Few-Shot Learners / T. Brown [и др.] // Advances in Neural Information Processing Systems. т. 33 / под ред. H. Larochelle [и др.]. — Curran Associates, Inc., 2020. — с. 1877—1901. — URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.
9. Training language models to follow instructions with human feedback / L. Ouyang [и др.] // arXiv preprint arXiv:2203.02155. — 2022.
10. *OpenAI*. GPT-4 Technical Report. — 2023. — arXiv: 2303.08774 [cs.CL].
11. *SberDevices*. Русскоязычная нейросеть от Сбера. — доступ 3 ноября 2023. — https://developers.sber.ru/portal/products/gigachat.
12. Language is not all you need: Aligning perception with language models / S. Huang [и др.] // arXiv preprint arXiv:2302.14045. — 2023.

13. *AIRI*. RUDOLPH: One Hyper-Tasking Transformer can be creative as DALL-E and GPT-3 and smart as CLIP. — 2022. — https://github.com/ai-forever/rudolph.

14. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model / B. Workshop [и др.]. — 2023. — arXiv: 2211.05100 [cs.CL].

15. *Mikhailov V.*, *Serikov O.*, *Artemova E.* Morph Call: Probing Morphosyntactic Content of Multilingual Transformers // Proceedings of the Third Workshop on Computational Typology and Multilingual NLP. — Online : Association for Computational Linguistics, 06.2021. — с. 97—121. — DOI: 10.18653/v1/2021.sigtyp-1.10. — URL: https://aclanthology.org/2021.sigtyp-1.10.

16. Universal and Independent: Multilingual Probing Framework for Exhaustive Model Interpretation and Evaluation / O. Serikov [и др.] // Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. — Abu Dhabi, United Arab Emirates (Hybrid) : Association for Computational Linguistics, 12.2022. — с. 441—456. — URL: https://aclanthology.org/2022.blackboxnlp-1.37.

17. SemSketches2021: Experimenting with the machine processing of the pilot semantic sketches corpus | SemSketches2021: опыт автоматической обработки пилотного корпуса семантических скетчей //. т. 20. Komp'juternaja Lingvistika i Intellektual'nye Tehnologii. — 2021. — с. 560—570.

18. Высокоуровневая семантическая интерпретация структуры статических моделей для русского языка / О. А. Сериков [и др.] // Вестник Новосибирского государственного университета. Серия: Лингвистика и межкультурная коммуникация. — 2023. — т. 21, № 1. — с. 67—82.

19. Attention Understands Semantic Relations / A. Chizhikova [и др.] // Proceedings of the Thirteenth Language Resources and Evaluation Conference. — Marseille, France : European Language Resources Association, 06.2022. — с. 4040—4050. — URL: https://aclanthology.org/2022.lrec-1.430.

20. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing / под ред. Y. Goldberg, Z. Kozareva, Y. Zhang. — Abu Dhabi, United Arab Emirates : Association for Computational Linguistics, 12.2022. — с. 11689—11698. — URL: https://aclanthology.org/2022.emnlp-main.803.

21. Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP / под ред. J. Bastings [и др.]. — Abu Dhabi, United Arab Emirates (Hybrid) : Association for Computational Linguistics, 12.2022. — URL: https://aclanthology.org/2022.blackboxnlp-1.0.

22. Emergent structures and training dynamics in large language models / R. Teehan [и др.] // Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models. — 2022. — с. 146—159.

23. Is neural language acquisition similar to natural? A chronological probing study | Усвоение языка у языковых моделей и человека: хронологическое пробинг-исследование //. т. 21. Komp'juternaja Lingvistika i Intellektual'nye Tehnologii. — 2022. — с. 550—563.

24. On the Opportunities and Risks of Foundation Models / R. Bommasani [и др.] // CoRR. — 2021. — т. abs/2108.07258. — arXiv: 2108.07258. — URL: https://arxiv.org/abs/2108.07258.

25. Self-Organization and Artificial Life / C. Gershenson [и др.] // Artificial Life. — 2020. — сент. — т. 26, № 3. — с. 391—408. — DOI: 10.1162/artl_a_00324. — eprint: https://direct.mit.edu/artl/article-pdf/26/3/391/1896088/artl_a_00324.pdf. — URL: https://doi.org/10.1162/artl%5C_a%5C_00324.

26. *Dresp-Langley B.* Seven Properties of Self-Organization in the Human Brain // Big Data and Cognitive Computing. — 2020. — т. 4, № 2. — DOI: 10.3390/bdcc4020010. — URL: https://www.mdpi.com/2504-2289/4/2/10.

27. *Gordon D. M.* The ecology of collective behavior in ants // Annual review of entomology. — 2019. — т. 64. — с. 35—50.

28. *Dalrymple M.* Lexical functional grammar. — Brill, 2001.

29. *McCoy R. T., Min J., Linzen T.* BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance // Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. — Online : Association for Computational Linguistics, 11.2020. — с. 217—227. — DOI: 10.18653/v1/2020.blackboxnlp-1.21. — URL: https://aclanthology.org/2020.blackboxnlp-1.21.

30. Probing for the Usage of Grammatical Number / K. Lasri [и др.] // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Dublin, Ireland : Association for Computational Linguistics, 05.2022. — с. 8818—8831. — DOI: 10.18653/v1/2022.acl-long.603. — URL: https://aclanthology.org/2022.acl-long.603.

31. An overview of early vision in inceptionv1 / C. Olah [и др.] // Distill. — 2020. — т. 5, № 4. — e00024—002.

32. What do you learn from context? Probing for sentence structure in contextualized word representations / I. Tenney [и др.] // arXiv e-prints. — 2019. — май. — arXiv:1905.06316. — arXiv: 1905.06316 [cs.CL].

33. Analyzing individual neurons in pre-trained language models / N. Durrani [и др.] // arXiv preprint arXiv:2010.02695. — 2020.

34. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties / A. Conneau [и др.] // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Melbourne, Australia : Association for Computational Linguistics, 07.2018. — с. 2126—2136. — DOI: 10.18653/v1/P18-1198. — URL: https://www.aclweb.org/anthology/P18-1198.

35. *Kitaev N.*, *Klein D.* Constituency Parsing with a Self-Attentive Encoder // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Melbourne, Australia : Association for Computational Linguistics, 07.2018. — с. 2676—2686. — DOI: 10.18653/v1/P18-1249. — URL: https://aclanthology.org/P18-1249.

36. Jointly Predicting Predicates and Arguments in Neural Semantic Role Labeling / L. He [и др.] // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). — Melbourne, Australia : Association for Computational Linguistics, 07.2018. — с. 364—369. — DOI: 10.18653/v1/P18-2058. — URL: https://aclanthology.org/P18-2058.

37. Linguistically-Informed Self-Attention for Semantic Role Labeling / E. Strubell [и др.] // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. — Brussels, Belgium : Association for Computational

Linguistics, 11.2018. — c. 5027—5038. — DOI: 10.18653/v1/D18-1548. — URL: https://aclanthology.org/D18-1548.

38. *Lee K.*, *He L.*, *Zettlemoyer L.* Higher-Order Coreference Resolution with Coarse-to-Fine Inference // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). — New Orleans, Louisiana : Association for Computational Linguistics, 06.2018. — c. 687—692. — DOI: 10.18653/v1/N18-2108. — URL: https://aclanthology.org/N18-2108.

39. What do Neural Machine Translation Models Learn about Morphology? / Y. Belinkov [и др.] // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Vancouver, Canada : Association for Computational Linguistics, 07.2017. — c. 861—872. — DOI: 10.18653/v1/P17-1080. — URL: https://aclanthology.org/P17-1080.

40. Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks / Y. Belinkov [и др.] // Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). — Taipei, Taiwan : Asian Federation of Natural Language Processing, 11.2017. — c. 1—10. — URL: https://aclanthology.org/I17-1001.

41. Dissecting Contextual Word Embeddings: Architecture and Representation / M. E. Peters [и др.] // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. — Brussels, Belgium : Association for Computational Linguistics, 10.2018. — c. 1499—1509. — DOI: 10.18653/v1/D18-1179. — URL: https://aclanthology.org/D18-1179.

42. *Zhang K.*, *Bowman S.* Language Modeling Teaches You More than Translation Does: Lessons Learned Through Auxiliary Syntactic Task Analysis // Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. — Brussels, Belgium : Association for Computational Linguistics, 11.2018. — c. 359—361. — DOI: 10.18653/v1/W18-5448. — URL: https://aclanthology.org/W18-5448.

43. *Alain G.*, *Bengio Y.* Understanding intermediate layers using linear classifier probes // 5th International Conference on Learning Representations, ICLR

2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings. — OpenReview.net, 2017. — URL: https://openreview.net/forum?id=HJ4-rAVtl.

44. *Hewitt J.*, *Manning C. D.* A Structural Probe for Finding Syntax in Word Representations // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). — Minneapolis, Minnesota : Association for Computational Linguistics, 06.2019. — с. 4129—4138. — DOI: 10.18653/v1/N19-1419. — URL: https://aclanthology.org/N19-1419.

45. *Hewitt J.*, *Liang P.* Designing and Interpreting Probes with Control Tasks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). — Hong Kong, China : Association for Computational Linguistics, 11.2019. — с. 2733—2743. — DOI: 10.18653/v1/D19-1275. — URL: https://aclanthology.org/D19-1275.

46. *Chi E. A.*, *Hewitt J.*, *Manning C. D.* Finding Universal Grammatical Relations in Multilingual BERT // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — Online : Association for Computational Linguistics, 07.2020. — с. 5564—5577. — DOI: 10.18653/v1/2020.acl-main.493. — URL: https://aclanthology.org/2020.acl-main.493.

47. Linspector: Multilingual probing tasks for word representations / G. G. Şahin [и др.] // Computational Linguistics. — 2020. — т. 46, № 2. — с. 335—385.

48. Probing for Semantic Classes: Diagnosing the Meaning Content of Word Embeddings / Y. Yaghoobzadeh [и др.] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 07.2019. — с. 5740—5753. — DOI: 10.18653/v1/P19-1574. — URL: https://aclanthology.org/P19-1574.

49. *Rogers A.*, *Kovaleva O.*, *Rumshisky A.* A primer in bertology: What we know about how bert works // Transactions of the Association for Computational Linguistics. — 2020. — т. 8. — с. 842—866.

50. *Nie A.*, *Bennett E.*, *Goodman N.* DisSent: Learning sentence representations from explicit discourse relations // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — 2019. — с. 4497—4510.

51. *Chen M.*, *Chu Z.*, *Gimpel K.* Evaluation benchmarks and learning criteria for discourse-aware sentence representations // arXiv preprint arXiv:1909.00142. — 2019.

52. BLiMP: The benchmark of linguistic minimal pairs for English / A. Warstadt [и др.] // Transactions of the Association for Computational Linguistics. — 2020. — т. 8. — с. 377—392.

53. *Kuratov Y.*, *Arkhipov M.* Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. — 2019. — arXiv: 1905.07213 [cs.CL].

54. *Straka M.*, *Straková J.* Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe // Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. — 2017. — с. 88—99.

55. *Loureiro D.*, *Jorge A.* Language Modelling Makes Sense: Propagating Representations through WordNet for Full-Coverage Word Sense Disambiguation // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 07.2019. — с. 5682—5691. — DOI: 10.18653/v1/P19-1569. — URL: https://aclanthology.org/P19-1569.

56. *Vrandečić D.*, *Krötzsch M.* Wikidata: a free collaborative knowledgebase // Communications of the ACM. — 2014. — т. 57, № 10. — с. 78—85.