Федеральное государственное автономное образовательное учреждение
высшего образования «Национальный исследовательский университет
«Высшая школа экономики»

*На правах рукописи*

**Гончарова Елизавета Федоровна**

# Выявление релевантных дискурсивных признаков для решения задач анализа текстовых данных

РЕЗЮМЕ
диссертации на соискание учёной степени
кандидата компьютерных наук

Москва - 2022

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

Научный руководитель: Сергей Олегович Кузнецов, д.ф.-м.н., профессор, Национальный исследовательский университет «Высшая школа экономики»

# ТЕМА ИССЛЕДОВАНИЯ

**Актуальность темы исследования.** Последние годы область обработки естественного языка (NLP) непосредственно связана с большими предобученными языковыми моделями, основанными на архитектуре Трансформер [49]. Данные языковые модели, такие как BERT [12], T5 [38] или серия моделей GPT [37], показали значительный прирост качества решения ряда текстовых задач, обогнав классические модели машинного обучения, а также все существующие до этого нейросетевые подходы. По сравнению с более ранними методами языкового моделирования [32], новые модели способны конструировать контекстуальные векторные представления слов, что, в свою очередь, позволяет им учитывать больше значимой информации о тексте и, в целом, о языке.

Несмотря на успех данного класса моделей, исследователи задаются вопросом, какую лингвистическую информацию о языке они способны кодировать в конструируемое векторное представление. В ряде работ было показано, что некоторые типы лингвистической информации действительно кодируются отдельными слоями сети или их комбинациями [42], однако, не все уровни языка одинаковым образом представлены в скрытых векторных представлениях, конструируемых моделью, что негативно влияет на качество решения сложных задач обработки естественного языка [59]. Например, модели на базе архитектуры Трансформер, зачастую не улавливают зависимость между частями текстов, которые находятся далеко друг от друга, что может быть критически важным при решении таких задач, как реферирование и симплификация текстов, понимание прочитанного текста или их генерация. В целом, проведенные исследования показывают, что высокие результаты предобученных языковых моделей на бенчмарках, связанных с оценкой степени понимания естественного языка обусловлены, в первую очередь, большим количеством обучаемых параметров нейронной сети, а также большим объемом текстовых данных, на которых были обучены модели, а не их способностью понимать лингвистическую структуру языка [33]. Несмотря на то, что для некоторых задач, эта особенность модели не является критичной (например, для задачи анализа тональности текстов), существует ряд задач, которые значительным образом зависят от степени понимания моделью этой структуры. Например, анализ

аргументированности текстов непосредственно зависит от способности модели понимать дискурс; для более точного нахождения ответов на вопросы, заданные к тексту, модели необходимо одновременно анализировать синтаксис, семантику и дискурс.

На настоящий момент, актуальной задачей является разработка языковых моделей, которые способны учитывать сложные лингвистические признаки при решении задач обработки естественного языка. В существующих подходах к решению данной задачи, авторы добавляют дополнительные знания о языке в предобученную модель за счет обучения дополнительного нейросетевого компонента, отвечающего за представление лингвистической структуры [57, 55].

В данной работе исследуется способность предобученных языковых моделей на базе архитектуры Трансформер анализировать дискурсивную структуру, которая, в отличие от синтаксиса, определяется не на уровне предложения, а на уровне всего текста и выражает логическую организацию мыслей автора [17] текста. В работе также показывается, что анализ дискурсивной структуры является важным для ряда прикладных задач обработки естественного языка, для решения которых модели необходимо учитывать зависимость между различными частями текста, а не только анализировать контекст слов.

Ключевым аспектом при решении данной задачи является выбор подхода по кодированию дискурсивной информации в языковую модель. Существующие работы, в основном, применяют графовые нейронные сети для конструирования векторного представления дискурсивной структуры, а затем объединяют полученные представления с контекстуальными векторами, построенными для текста языковой моделью. Подобный подход был успешно применен для решения задачи реферирования текста [53]. Однако, обучение подобного рода моделей-ансамблей требует большого количества размеченных текстов, а также увеличивает вычислительные затраты на обучение модели. В данной работе предлагается новый подход, позволяющий предобученной модели учитывать дискурсивную структуру без внесения существенных изменений в архитектуру модели. В работе также исследуется влияние дискурсивных признаков на решение ряда задач обработки естественного языка и показывается, что учет

дискурсивной структуры положительно влияет на качество решения задач оценки аргументированности текста, понимания прочитанного текста, а также задачи интерпретации результатов работы языковой модели.

**Цель и задачи исследования.**

Таким образом, **целью** данного исследования является оценка значимости дискурсивных признаков для предобученных языковых моделей, а также разработка новых подходов по кодированию дискурсивной информации в данные модели, которые позволят улучшить качество решения прикладных задач обработки естественного языка.

Для достижения цели проводимого исследования в работе был поставлен ряд **задач**:

1. Оценить значимость дискурсивной структуры текста за счет оценки качества решения сложных текстовых задач существующими языковыми моделями, не учитывающими дискурс.
2. Предложить и реализовать новый подход для кодирования дискурсивной структуры предобученной моделью BERT, который позволит ей кодировать дискурс в конструируемые векторные представления. Проверить качество работы модели при решении задачи оценки аргументированности текстов.
3. Предложить и реализовать новый подход для создания дискурсивно-обогащенного механизма внимания, встроенного в предобученную модель BERT. Проверить качество работы модели при решении задачи понимания прочитанного текста.
4. Разработать и реализовать новый метод по извлечению *текстовых обоснований*, служащих для интерпретации и объяснения результатов работы языковой модели, основанный на дискурсивном анализе текстов.

## ОСНОВНЫЕ РЕЗУЛЬТАТЫ

**Научная новизна.** В данной работе впервые установлены следующие положения.

1. Выполнен анализ способности предобученных языковых моделей учитывать дискурсивную структуру, проанализированы существующие подходы к кодированию дискурса в языковые модели.

2. Разработан и реализован новый подход к кодированию дискурсивной структуры в предобученную языковую модель BERT (*disBERT*), для обучения модели предложена новая задача дискурсивно-обусловленного маскированного языкового моделирования.
3. Предложена схема перевода дискурсивного дерева разбора в граф дискурсивных зависимостей, позволяющий сохранять только информативные дискурсивные признаки.
4. Разработан и реализован подход по внедрению дискурса в механизм внимания, что позволило модели находить более релевантные фрагменты текста при решении задачи понимания прочитанного текста.
5. Предложен новый подход для интерпретации результатов работы моделей на базе архитектуры Трансформер. Разработанный метод объединяет интерпретируемые модели машинного обучения и нейросетевые модели для выявления *текстовых объяснений*.

**Практическая значимость.** К практическим результатам данного исследования можно отнести следующее:

1. Разработанные модели могут применяться для создания дискурсивно-обогащенных контекстуальных векторных представлений, которые могут применяться для решения прикладных задач обработки естественного языка.
2. Разработанная модель для интерпретации результатов позволяет генерировать *текстовые объяснения* при решении задач обработки естественного языка с помощью предобученных языковых моделей.
3. Модель disBERT была применена для фильтрации пользовательских отзывов о товарах при разработке чатбота для электронной коммерции [30].
4. Код разработанных моделей и размеченные наборы данных выложены в открытый доступ[1] и могут применяться исследователями для решения задач, для которых необходимо учитывание дискурса.

[1] https://github.com/lizagonch/Discourse-BERT

**Методология и методы исследования.** В работе использованы методы машинного обучения, в частности, глубинного обучения, теории вероятностей и статистики, дискурсивного анализа, а также анализа формальных понятий. Код реализации моделей написан на языке Python с использованием фреймворка pytorch для работы с моделями глубинного обучения, для имплементации моделей на базе архитектуры Трансформер использовалась библиотека transformers от Hugging Face, в коде также применяются библиотеки numpy и sklearn, которые активно используются в сообществе при решении схожих задач, что удовлетворяет принципам воспроизводимости результатов.

**Основные положения, выносимые на защиту:**

1. Предложена и обучена модель, которая учитывает дискурсивную структуру текста за счет ее кодирования во входной слой модели и дообучения с помощью дискурсивно-обогащенной задачи маскированного языкового моделирования.
2. Предложена и обучена модель, обладающая дополнительным слоем внимания, зависимого от дискурсивной структуры текста.
3. Предложен алгоритм по переводу дискурсивного дерева разбора в дискурсивный граф зависимостей, в котором можно учитывать только релевантные для решаемой задачи дискурсивные признаки.
4. Предложен метод по конструированию *текстовых обоснований* для объяснений и интерпретации ответов модели на основании анализа дискурсивной структуры.

**Личный вклад в положения, выносимые на защиту.** Дискурсивно-обогащенная модель disBERT и модифицированная задача маскированного языкового моделирования разработана и реализована автором лично. Алгоритм для перевода дискурсивного дерева разбора в граф дискурсивных зависимостей с учетом релевантных признаков также является разработкой автора диссертации. Модель BERT, обогащенная дискурсивным механизмом внимания, разработана автором диссертации совместно с соавторами Б.А. Галицким и Д.А. Ильвовским. Автором диссертации было предложено использвать информацию из графа дискурсивных зависимостей для механизма внимания, а также был проведен эксперимент по получению векторных представлений для каждой из дискурсивных единиц текста

по отдельности. Автором диссертации также был произведен сбор и дискурсивная разметка наборов данных, используемых в экспериментах. Метод для конструирования *текстовых обоснований* является собственной разработкой автора, при этом, используемый алгоритм по нахождению релевантных дискурсивных схем с помощью анализа формальных понятий был разработан совместно с научным руководителем С.О. Кузнецовым.

## ПУБЛИКАЦИИ И АПРОБАЦИЯ РАБОТЫ

**Апробация работы.** Основные результаты проведенного исследования были представлены на ряде конференций в области интеллектуального анализа данных и обработки естественного языка, а также опубликованы в 6 работах, проиндексированных в базах данных Scopus и Web of Science.

**Публикации стандартного уровня.**

1. Goncharova E.[1] Relying on Discourse Analysis to Answer Complex Questions by Neural Machine Reading Comprehension (Применение дискурсивного анализа для ответа на сложные вопросы при решении задачи понимания прочитанного текста нейросетевыми моделями) / Galitsky B., Ilvovsky D., Goncharova E. // Proceedings of the International Conference Recent Advances in Natural Language Processing. С. 444-453, 2021. (Scopus)
2. Goncharova E. Relying on Discourse Trees to Extract Medical Ontologies from Text (Использование дискурсивных деревьев для конструирования онтологий по медицинским текстам) / Galitsky B., Ilvovsky D., Goncharova E. // Lecture Notes in Computer Science, Том 12948 LNAI, С. 215 – 231, 2021. (Scopus – Q2)
3. Goncharova E.[1] Concept-based chatbot for interactive query refinement in product search (Разработка чат-бота с применением анализа формальных понятий для интерактивного уточнения пользовательских запросов при поиске товаров) / Goncharova, E., Ilvovsky, D., Galitsky, B. // CEUR Workshop Proceedings, Том 2972, С. 51 – 58, 2021. (Scopus)

[1]Автор диссертации является главным автором работы.

4. Goncharova, E. On a chatbot conducting dialogue-in-dialogue (демо-статья) (Разработка чат-бота, генерирующего диалог из повествовательного текста (диалог-в-диалоге)) / Galitsky, B., Ilvovsky, D., Goncharova, E. // SIGDIAL 2019 - 20th Annual Meeting of the Special Interest Group Discourse Dialogue - Proceedings of the Conference, С. 118 – 121, 2019. (Scopus, Web of Science)
5. Goncharova E. On a Chatbot Providing Virtual Dialogues (Разработка чат-бота, способного вести виртуальный диалог) / Galitsky B., Ilvovsky D., Goncharova E. // International Conference Recent Advances in Natural Language Processing, RANLP, Том 2019-September, С. 382 - 387, Код 155296, 2019. (Scopus)
6. Goncharova E.[1] Increasing the efficiency of packet classifiers with closed descriptions (Повышение качества классификации пакетов данных за счет применения анализа формальных понятий) / Goncharova E.F., Kuznetsov S.O. // CEUR Workshop Proceedings, Том 2529, С. 75 – 88, 2019. (Scopus)

**Доклады на конференциях и семинарах.**

1. International Conference Recent Advances in Natural Language Processing (RANLP 2021), доклад: *Relying on Discourse Analysis to Answer Complex Questions by Neural Machine Reading Comprehension*, 1-3 сентября 2021, Варна, Болгария (онлайн).
2. COLING Workshop on Natural Language Processing in E-Commerce (EComNLP 2020), доклад: *On a Chatbot Navigating a User through a Concept-Based Knowledge Model*, 12 декабря 2020, Барселона, Испания (онлайн).
3. 18th Russian Conference on Artificial Intelligence (RCAI-2020), доклад: *FCA-based Approach for Interactive Query Refinement with IR-chatbots*, 10-16 октября 2020, Москва, Россия.
4. The 20th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2019), доклад (демо трек): *On a Chatbot Conducting Dialogue-in-Dialogue*, 11-13 сентября 2019, Стокгольм, Швеция.

5. The 3rd International Workshop Formal Concept Analysis for Knowledge Discovery (FCA4KD), доклад: *Increasing the efficiency of packet classifiers based on closed descriptions*, 7 июня 2019, Москва, Россия.

## СОДЕРЖАНИЕ РАБОТЫ

Во **введении** диссертации описывается актуальность проведенного исследования, ставятся цели и задачи исследования, а также описываются ключевые результаты, полученные в рамках выполнения работы. В данной главе также кратко затрагиваются результаты предыдущих работ по данной тематике и приводятся основные научные задачи, которые еще не были решены исследователями в данной области, и на которые обращено внимание в данной работе.

В **первой главе** приводится описание основных эпох развития языковых моделей, начиная от статистических подходов и заканчивая глубокими нейронными сетями, реализующими архитектуру Трансформер. Целью современных языковых моделей является конструирование векторного представления текста (эмбеддинга), в котором может быть закодирована вся доступная и релевантная для решаемой задачи лингвистическая информация.

Так как основная часть диссертация посвящена анализу предобученных языковых моделей, построенных на базе архитектуры Трансформер, языковое моделирование (LM-ing) рассматривается с точки зрения базовой процедуры для предобучения данного типа моделей. Как правило, языковая модель предсказывает следующее слово $w_N$ в последовательности учитывая его левый контекст: $[w_1,w_2,...,w_{N-1}]$. Таким образом, целью вероятностного языкового моделирования является либо вычисление вероятности последовательности слов: $P(w_1,w_2,...,w_{N-1},w_N)$, либо нахождение вероятности следующего слова в последовательности: $P(w_N|w_1,w_2,...,w_{N-1})$. При генерации нового слова языковая модель оценивает вероятность появления каждого слова, присутствующего в словаре.

Маскированное языковое моделирование (MLM) – это модификация стандартной процедуры языкового моделирования, которая также может применяться для предобучения больших языковых моделей, основанных на

архитектуре Трансформер, таких как BERT и его вариации. В MLM $i$-ое слово во входной последовательности скрыто, и модель учится вычислять распределение вероятности для данного скрытого слова по открытому левому и правому контексту, $P(w_i|w_1,...,w_{i-1},w_{i+1},...,w_N)$, где $N$ – количество слов во входной последовательности.

Для предобучения моделей, основанных на архитектуре Трансформер, применяются задачи языкового или маскированного языкового моделирования совместно с большими коллекциями неразмеченных текстовых корпусов на естественном языке. Эта комбинация позволяет обучить достаточно успешные языковые модели, которые обладают некоторым пониманием естественного языка, а также успешно решают ряд языковых задач после дообучения либо в так называемом zero-shot режиме, т.е. без обучающих примеров.

В настоящее время, подходы, основанные на предобучении моделей на больших корпусах неразмеченных текстов, полностью заменили классические методы машинного обучения в области обработки естественного языка, где последние строились с помощью выбранных и сгенерированных вручную лингвистических признаков. Для конструирования таких информативных признаков требовалось много ресурсов, связанных с привлечением профессиональных лингвистов, которые могли определить какая комбинация признаков лучшим образом справится с той или иной языковой задачей. Большим языковым моделям не требуется подобный подход к построению признаков. Они способны работать с так называемыми сырыми данными, автоматически конструируя релевантные числовые описание (скрытые векторные представления) на каждом слое нейронной сети, генерируя их в соответствии с решаемой задачей языкового моделирования.

Несмотря на успех нейросетевых моделей у них существует ряд недостатков, которые отмечают исследователи: во-первых, векторные представления, конструируемые языковыми моделями, зачастую не отражают лингвистические особенности языка. Таким образом, при решении ряда сложных задач (например, генерации текста, реферирования текста или ответа на вопросы) ответы, получаемые с помощью языковых моделей, могут быть грамматически правильными, но нелогичными или не

соответствовать нашим знаниям о мире. Кроме того, конструируемые векторные представления являются плохо интерпретируемыми, что ограничивает их применимость в таких важных областях, как юриспруденция или медицина.

В данном исследовании мы обращаемся к этим проблемам, предлагая подходы по включению релевантных лингвистических признаков в предобученные языковые модели для улучшения качества решения ряда сложных языковых задач, а также повышения интерпретируемости данных моделей.

В первой главе также более подробно описывается модель BERT, которая применяется в исследовании как базовая модель, основанная на архитектуре Трансформер, и для которой предлагаются модификации для повышения степени представленности дискурсивной информации в конструируемых векторных представлениях. BERT (Bidirectional Encoder Representations from Transformers) – это модель-кодировщик, состоящая из 12 слоев внимания (self-attention layers), которая способна создавать контекстуализированные векторные представления слов (или токенов), а также целых предложений. Классическая модель BERT обучалась с использованием двух задач: MLM и определения взаимосвязи между двумя предложениями во входной паре (Next sentence prediction, NSP). в задаче NSP было предложено предварительно обучать совместное представление пары предложений, предсказывая, насколько вероятно то, что два предложения могут следовать в тексте друг за другом. Исходная модель была предобучена на больших массивах немаркированных текстов, а именно BooksCorpus (800 млн. слов) [60] и английской Википедии (2500 млн. слов). Подобная вариативность задач, используемых для обучения, требует некоторых архитектурных особенностей, реализованных в модели, которые используются в представленном исследовании для кодирования дополнительной лингвистической информации.

Благодаря тому, что модель дополнительно обучена решать задачу NSP, она имеет специальный обучаемый входной слой, отвечающий за кодирование сегмента текста, который определяет, какие токены относятся к первому, а какие – ко второму предложению в паре. Мы предлагаем дообучать данный входной слой, который будет кодировать информацию

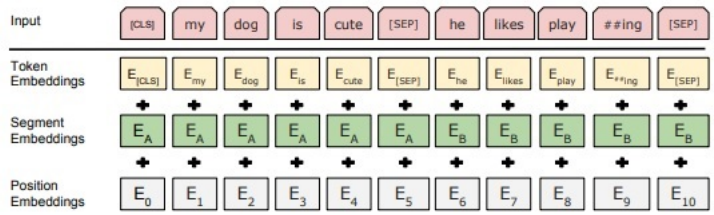


Рисунок 1 — Формат входных данных для модели BERT, включающий слой для кодирования входных токенов, кодирования сегмента текста, а также кодирования позиции токена в тексте.

о дискурсивных отношениях, связывающих два согласованных текстовых фрагмента. На рисунке 1 приведена схема представления входных данных для модели BERT.

**Вторая глава** диссертации посвящена вопросу качества представления различных видов лингвистических характеристик текста предобученными языковыми моделями на базе архитектуры Трансформер. В этой главе рассматриваются методы анализа языковых моделей (probing tasks), которые были предложены для оценки их способности кодировать лингвистические свойства языка (синтаксис, семантику, дискурс). В заключение главы приводится анализ результатов, полученных исследователями, делается краткий вывод о том, какие типы лингвистической информации улавливаются языковыми моделями, а какие слабо кодируются в векторные представления.

Высокое качество, которые показывают предобученные языковые модели при решении ряда задач обработки естественного языка, побудило исследователей к детальному анализу данных моделей и оценке векторных представлений, которые они конструируют, с точки зрения представленности в них лингвистической информации о языке. Результаты экспериментальных исследований, показывающие, что модели на базе архитектуры Трансформер значительным образом превосходят все классические модели машинного обучения, а также другие модели глубинного обучения при решении задач, оценивающих степень понимания естественного языка (GLUE [50], SuperGLUE [51]), позволяют предположить, что данные модели должны обладать способностью улавливать лингвистические характеристики языка [9]. Например, для решения задачи логического вывода (Recognizing Tex-

tual Entailment, RTE) модель должна понимать синтаксическую взаимосвязь между частями текста.

В целом, лингвистические признаки сгруппированы лингвистами в пять основных категорий: фонемы, морфемы, лексемы, синтаксис и контекст, которые, соответственно, изучаются такими разделами лингвистики, как: фонетика, фонология, морфология, синтаксис, семантика, дискурс и прагматика [4, 3]. Авторами ряда работ были предложены методы пробинга, которые служат для установления лингвистической осведомленности моделей по каждой из указанных лингвистических категорий.

К таким задачам можно отнести, например, разметку частей речи [43, 1, 5], устранения неоднозначности смысла слов [35]. В работе [47] авторы предложили общую архитектуру задач пробинга, которые могут использоваться для проверки разного типа лингвистической осведомленности моделей. В общем случае, задачи пробинга языковых моделей – это вспомогательные задачи, оценивающие конструируемые моделями скрытые представления текстов на разнообразных языковых задачах, требующих знание того или иного лингвистического признака.

После проверки качества моделей при решении различных задач пробинга авторами был сделан вывод, что, в большинстве случаев, механизм внимания, используемый в моделях, основанных на архитектуре Трансформер, позволяет им улавливать лингвистические признаки на уровне слов и предложений, т.е. семантику и синтаксис; более сложные функции на уровне текста, такие, например, как дискурс, достаточно слабо отражены в конструируемых векторных представлениях.

В **третьей главе** рассматривается дискурсивная структура текста и теория риторических структур (ТРС), используемая в данной работе для описания дискурса. В данной главе также анализируются вариации задач для дискурсивного пробинга и приводится обзор популярных подходов по включению дискурсивной информации в векторные представления, конструируемые языковыми моделями.

Дискурсивная структура описывает взаимосвязи между отдельными текстовыми фрагментами, а не между частями одного предложения как синтаксическая структура [23]. Данная структура описывает весь текст в виде согласованных дискурсивных единиц, связанных определенным

семантическим отношением. Можно отметить, что эта структура отражает логическую организацию мыслей автора, развивающуюся на протяжении всего текста.

Дискурсивная структура текста может быть описана с помощью различных фреймворков таких как TPC [31], PDTB [36] или Graph Bank [52]. TPC – один из самых популярных фреймворков, используемых авторами при проведении дискурсивного анализа. Его популярность обусловлена доступностью дискурсивных парсеров, обеспечивающих хорошее качество при дискурсивной разметке, а также его способностью строить дискурсивную структуру как комбинацию семантических и интенциональных отношений, которая охватывает широкий спектр логических связей, которые могут существовать между текстовыми фрагментами [29].

В данном исследовании мы также применяем TPC [31] для описания дискурсивной структуры. Согласно данной теории, текст может быть разделен на непересекающиеся текстовые фрагменты – элементарные дискурсивные единицы (ЭДЕ). Каждая ЭДЕ может выступать в качестве ядра или спутника в дискурсивном отношении, при этом, ядра являются главным элементом в отношении, а спутники – дополнительным. Ядра заключают смысловую информацию, которую автор выражает в тексте, а спутники содержат дополнительную, или вспомогательную, информацию, подтверждающую ту, которая представлена в ядре. В результате дискурсивная структура текста может быть представлена в виде иерархически-организованного дискурсивного дерева (ДД). Пример ДД представлен на рисунке 3 (b).

В PCT выделяют две основные категории дискурсивных отношений: многоядерные (N-N) отношения и отношения вида ядро-спутник (N-S), которые также называют антисимметричными. В антисимметричных отношениях, которые могут быть установлены в паре ЭДЕ, ядро является главной частью отношения, а спутник – подчиненной частью, которая зависит от ядра.

Исследования показали, что учет различий между многоядерными и антисимметричными отношениями может быть полезен при решении ряда прикладных задач области обработки естественного языка, например, при разрешении анафоры [11, 20] или при создании вопросно-ответных систем.

При решении задачи понимания прочитанного текста, для нахождения релевантного текстового фрагмента, отвечающего на поставленный вопрос, может быть полезно учитывать дискурсивную структуру текста [15]. Так, например, нами было проведено исследование и выявлены некоторые зависимости между типами дискурсивных отношений и их соответствия различным типам специальных вопросов, начинающихся со слов *wh-*, [13, 14, 16]. Например, N-S отношение *Attribution* соответствует вопросам *what/who is the source*, где ответ находится в ядре найденного отношения; спутники отношений *Cause* и *Explanation* могут служить ответами на вопросы *why ...*. Обладая этой информацией, модель сможет максимально точно связывать различные виды вопросов с релевантными фрагментами текста, в которых может находиться ответ на каждый конкретный вопрос.

Задачи дискурсивного пробинга были предложены в ряде работ последних лет. Одно из самых полных исследований по изучению дискурсивной осведомленности предобученных моделей приводится в статье [26]. Авторы работы предлагают многоуровневый механизм проверки кодирования дискурсивной информации предобученными моделями. Авторы также отмечают, что существующие методы пробинга направлены, в основном, на анализ степени выявления синтаксических связей в тексте, в то время как дискурсивной структура часто остается за рамками анализа. В работе было предложено семь задач, для решения которых модели необходимо обладать информацией о дискурсивной структуре текста. После проверки наиболее известных предобученных языковых моделей, авторы сделали выводы о том, что базовая модель BERT хорошо справляется с задачами на уровне определения взаимосвязи между двумя предложениями (что связано с ее предобучением на подобном типе задач), в то время как на дискурсивных задачах, например, задача предсказания дискурсивного маркера или определения типа дискурсивного отношения, модель показывает слабые результаты.

В нашем исследовании мы используем одну из предложенных задач пробинга – *предсказание дискурсивного маркера* – для проверки способности улучшенной языковой модели BERT понимать дискурс. Под дискурсивными маркерами понимаются языковые единицы, которые характеризуют определенные дискурсивные отношения в тексте, а также

обеспечивают его связность и целостность. К дискурсивным маркерам можно отнести, например, союзы (*but*, *because*, *and*), наречия (*then*, *so*) и другие.

В **четвертой главе** данной диссертации описывается модель *disBERT*, которая является расширением оригинальной модели BERT и способна учитывать дискурсивную организацию текста. Для обеспечения модели дополнительной дискурсивной информацией была предложена модифицированная задача условного маскированного языкового моделирования, благодаря которой модель обучается предсказывать скрытое слово не только по его контексту, но и по типу дискурсивного отношения, соответствующего рассматриваемому текстовому фрагменту. В данной главе, мы показываем, что для правильного решения стандартной задачи MLM, модели, в некоторых случаях, может не хватать информации о дискурсивной организации текста. Исходя из этого мы предлагаем расширение существующей задачи MLM за счет добавления в соответствующую ей функцию потерь дополнительного условия, характеризующегося дискурсивными признаками.

Рассмотрим простой пример ситуации, когда предобученная модель оказывается неспособна правильно предсказать замаскированный токен только по его контексту.

*He went out* ***while*** *it was not raining. [Condition]*
*He went out* ***and*** *it was not raining. [Elaboration]*

Если заменить слова, выделенные полужирным начертанием, на специальный токен-маску, то предобученная модель BERT оказывается неспособна правильно предсказать, какое именно слово скрыто маской, без дополнительной информации о его дискурсивной роли в предложении.

Для данного примера, мы запустили предобученную модель BERT[2] для предсказания замаскированного слова в предложении *'He went out [MACKA] dit was not raining'*. Модель выдала пять наиболее вероятных предсказаний: *"when"*, *"and"*, *точка с запятой*, *запятая* и *"but"*. Предобученная модель BERT приняла решение на основе примеров, которые она видела в обучающем корпусе, в то время как правильный ответ должен зависеть не только от контекста, но и от того, какую дискурсивную роль

[2]https://demo.allennlp.org/masked-lm

выполняет замаскированное слово в тексте. Несмотря на то, что предложение с *"when"* является грамматически и логически правильным, оно меняет смысл исходных предложений.

В данной работе предлагается задача маскированного языкового моделирования, обусловленная дискурсивной структурой, которую можно формализовать следующим образом. Мы обучаем модель предсказывать скрытое слово по контексту, окружающему данное слово, а также по определенному дискурсивному отношению, которое соответствует данному контексту. При такой постановке задачи, модель учится предсказывать $p\left(\cdot|C_{\backslash w_t}, y_t\right)$, где $y_t$ – метка, соответствующая дискурсивному отношению, вместо $p\left(\cdot|C_{\backslash w_t}\right)$. При обучении модели для такой задач мы минимизируем соответствующую функцию потерь

$$L_{MLM}(D_I|D\backslash\{D_I\}) = \frac{1}{K}\sum_{k=1}^{K}\log p(w_{i_k}|D\backslash\{D_I\}; y_{i_k}; \theta),$$

где $D_I$ – множество скрытых токенов входной последовательности, $w_{i_k}$ – предсказанный токен, $K$ – количество замаскированных токенов, $y_{i_k}$ – дискурсивное отношение, соответствующее замаскированному токену $w_{i_k}$, а $\theta$ – это параметры модели.

Для того чтобы обеспечить кодирование дискурса, было предложено трансформировать входной слой модели BERT, кодирующий сегмент предложения в задаче NSP, в *дискурсивный слой*, который будет кодировать соответствующее дискурсивное отношение в процессе обучения на модифицированной задаче маскированного языкового моделирования. Отметим, что сегментационный слой исходной модели BERT обучался по бинарным меткам, которые обозначали принадлежность токена первому или второму предложению в паре. В предлагаемой модифицированной схеме сегментационный слой должен кодировать более двух меток, что требует его переобучения до размера, совместимого с количеством дискурсивных отношений, которые могут быть обнаружены в тексте. Соответствующий сегментационный слой переобучается с нуля. Архитектура предложенной дискурсивной модели BERT (disBERT) приведена на рисунке 2.

**Представление данных в модели disBERT** Предложенная модель disBERT должна обрабатывать исходную последовательностью токенов, расширенную дополнительной дискурсивной информацией, извлеченной из дискурсивного дерева разбора. Однако иерархическая организация деревьев разбора не позволяет модели учитывать признаки, непосредственно извлеченные из этих деревьев. Таким образом, для запуска модели disBERT, дискурсивные признаки должны подаваться модели в соответствующем формате. В диссертации предложен алгоритм для представления дискурсивного дерева разбора в формате списка троек, где релевантное дискурсивное отношение соединяет две ЭДЕ или подмножество нескольких ЭДЕ. Для получения данных троек, дерево разбора должно быть переведено в граф дискурсивных зависимостей (ГДЗ), непосредственно из которого можно выделить тройки.

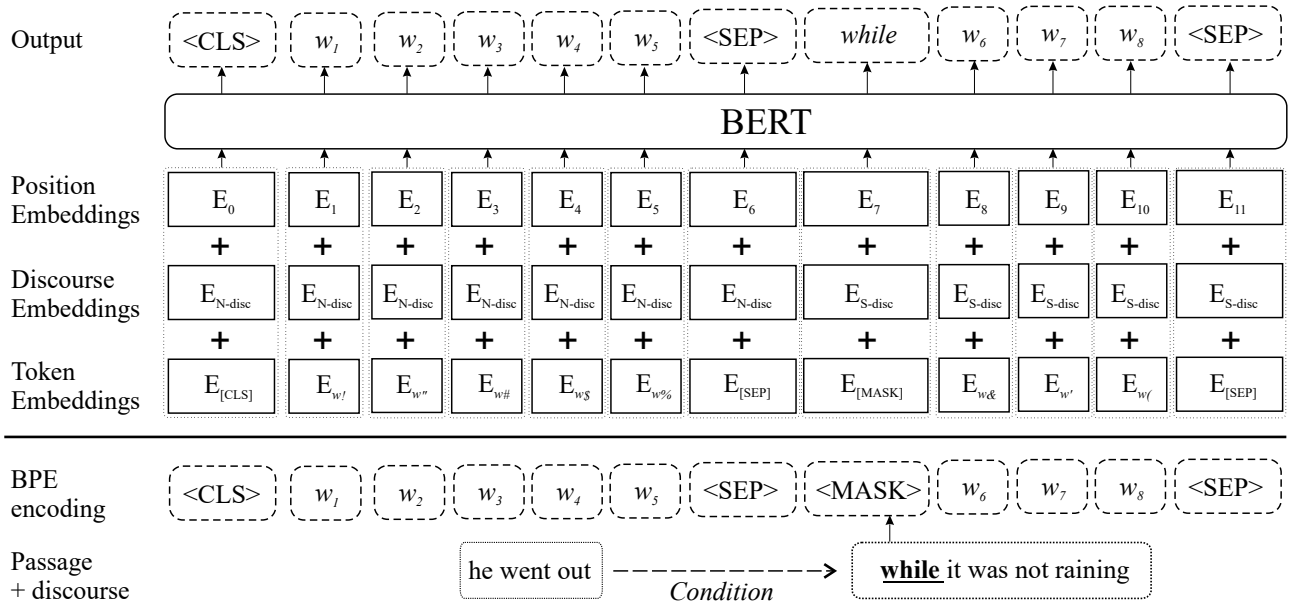


Рисунок 2 — Архитектура модифицированной модели disBERT, обогащенной дискурсом. На рисунке представлена модифицированная модель BERT, дополненная структурой дискурса. *Position Embeddings*, *Token Embeddings* и *Byte-Pair Encoding* являются стандартными слоями оригинального BERT. Напротив, слой *Discourse Embeddings* представляет собой модифицированный слой, колирующий тип дискурсивного отношения, которое связывает две ЭДЕ, разделенные специальными токенами [SEP] на входе модели disBERT.

Построение графа зависимостей начинается с пустого графа, затем мы обходим дискурсивное дерево снизу вверх, чтобы получить так называемые

главные вершины графа, соответствующие каждой вершине дерева (вершина дерева характеризует либо одну ЭДЕ, либо одно дискурсивное отношение). Далее, найденные главные вершины должны быть объединены ребром, соответствующим дискурсивному отношению, которое связывает две главные вершины в ДД. Таким образом, главная вершина выбирается в соответствии со следующими правилами:

– Главной вершиной для терминальной вершины дерева является сама терминальная вершина.
– Если у нетерминальной вершины есть дочерние элементы ядро и спутник, то главной вершиной назначается дочерний элемент – ядро, поскольку спутники соответствуют вспомогательным частям текста, зависящим от ядер.
– Если оба дочерних элемента нетерминальной вершины являются ядрами, то главной вершиной для него будет являться главная вершина его левого ребенка. Выбирается левый ребенок, потому что исходный текст анализируется слева направо.

Пример преобразования ДД → ГДЗ представлен на рисунке 3.

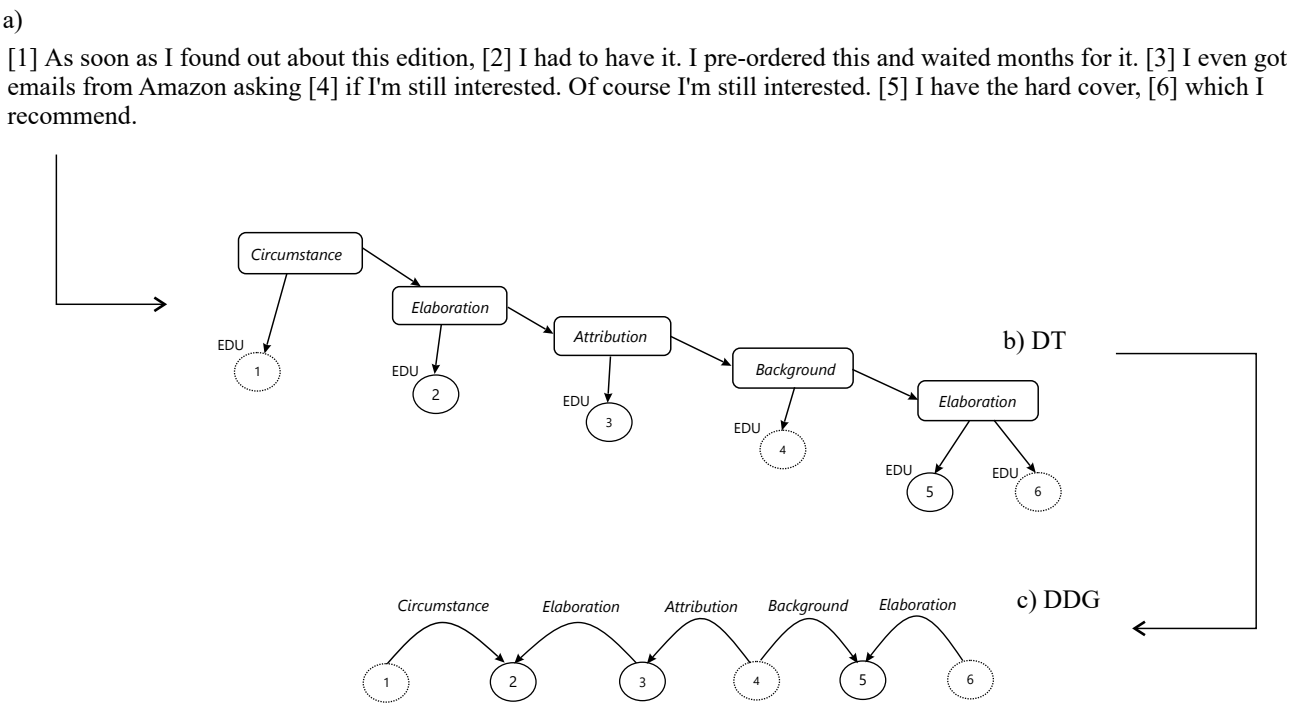


Рисунок 3 — Схема построения ГДЗ из ДД. Текст разбивается на 6 ЭДЕ (a), по которым строится ДД (b). Полученное ДД затем преобразуется в ГДЗ (c). Сплошные линии обозначают вершины-ядра, пунктирные линии определяют спутники. Дискурсивные отношения выделены курсивом.

**Результаты экспериментов.** Для оценки качества предложенной модели disBERT с точки зрения ее понимания дискурсивной структуры текста, мы запустили модель на одной из задач дискурсивного пробинга, предложенной в [26], а именно на задаче *предсказания дискурсивных маркеров*. Мы сравнили качество предсказания модели с результатами, приведенными авторами статьи для стандартных предобученных языковых моделей, и показали, что модель disBERT позволяет улучшить точность предсказания дискурсивных маркеров, в среднем, почти на 8%.

В качестве другого эксперимента, мы дообучили модель disBERT для решения задачи оценки аргументированности текстов (данная задача представлена в формате задачи бинарной классификации). В предыдущих работах было доказано, что для качественного решения задач, связанных с выявлением аргументационной структуры в тексте (argumentation mining, AM), модель машинного обучения должна обладать пониманием дискурсивной структуры текста, поэтому мы выбрали данную задачу для экспериментальной проверки качества работы предложенной модели disBERT.

При проведении эксперимента используется набор данных отзывов пользователей Amazon (AR dataset [34]) и UKP corpus [45]. В качестве базовых моделей были выбраны BiLSTM [45], MARGOT [28], MARGOT + TF-IDF [34], MARGOT + BoW, MARGOT + Discourse, $\text{BERT}_{base}$, BERT + Discourse [7], BERT-base, дообученные для решения рассматриваемой задачи.

Метрика $\text{F}_1$, точность (precision) и полнота (recall) использовались для численной оценки качества решения исследуемой задачи. В качестве дискурсивного парсера была применена модель, предложенная в работе [24][3], которая показывает одни из лучших результатов в дискурсивной разметке текстов на английском языке. Также в работе приведено исследование влияния дискурсивных парсеров на качество работы предложенной модели и выявлено, что полученные численные результаты незначительно изменяются при использовании разных парсеров.

Результаты работы моделей для двух наборов данных (AR и UKP) приведены в таблице 1. В таблице 2 показаны результаты модели disBERT для каждой из восьми тем из набора UKP. Этот эксперимент был проведен

[3]Модель доступна по ссылке https://github.com/jiyfeng/DPLP

для оценки зависимости представленной модели от предметной области, к которой относятся анализируемые тексты. В предыдущих исследованиях было показано, что при использовании стандартных предобученных языковых моделей качество решения задачи оценки аргументированности текстов значительно снижается в случаях, если модель не видела некоторые темы в процессе дообучения.

**Результаты для набора данных UKP Corpus.** Базовая модель BiLSTM показывает худшие результаты на наборе UKP по сравнению со всеми исследуемыми моделями. Стандартная модель $\text{BERT}_{base}$ значительно улучшает показатели всех метрик по сравнению с BiLSTM, при этом, стоит отметить, что полнота (recall), которая составляет 0,26 все еще довольно низкая для данной модели. Лингвистически-расширенные модели значительно улучшают качество решения поставленной задачи на наборе UKP. Так, например, учитывание дискурсивной информации моделью disBERT позволяет ей улучшить полноту примерно на 0,27 по сравнению со стандартной моделью $\text{BERT}_{base}$, при этом, она показывает результаты, сравнимые с результатами, полученными моделью BERT, расширенной информацией о теме, к которой относится входной текст. Модель, представленная в виде комбинации классификатора на основе BERT и XGBoost, обученного классификации текстов только на основании дискурсивных признаков, закодированных как one-hot векторы, незначительно превосходит качество модели $\text{BERT}_{base}$, однако показывает худшие результаты по сравнению с лингвистически-расширенными моделями.

В целом, результаты, полученные моделью disBERT немного лучше, чем у тематически-расширенной модели $\text{BERT}_{topic}$, специально предложенной для набора данных UKP corpus; улучшение составляет примерно 0,04 и 0,02 для точности (precision) и $\text{F}_1$-меры, соответственно.

**Результаты для набора данных AR Dataset.** Качество работы моделей на основе архитектуры Трансформер для набора отзывов пользователей AR значительно выше, чем для корпуса UKP. Набор данных AR включает в себя больше обучающих примеров, каждый из которых длиннее, чем текст из набора UKP Corpus, что является важным фактором для моделей, основанных на глубинном обучении. Стандартная модель $\text{BERT}_{base}$ превосходит модель MARGOT, разработанную для

| Набор данных | Модель | Precision | Recall | $F_1$ score |
|---|---|---|---|---|
| UKP | BiLSTM [45] | 0.41 | 0.16 | 0.23 |
| | $BERT_{base}$ | 0.55 | 0.26 | 0.35 |
| | BERT w. discourse [7] | **0.57** | 0.32 | 0.41 |
| | BERT-$base_{topic}$ [41] | 0.53 | 0.52 | 0.52 |
| | disBERT | 0.56 | **0.53** | **0.54** |
| "Movies and TV" | MARGOT [28] | 0.54 | 0.77 | 0.63 |
| | MARGOT (TF-IDF) [34] | 0.73 | 0.78 | 0.75 |
| | MARGOT w. BoW | 0.74 | **0.77** | 0.75 |
| | MARGOT w. disc. | 0.76 | 0.78 | 0.78 |
| | $BERT_{base}$ | 0.62 | 0.68 | 0.65 |
| | BERT w. discourse [7] | 0.65 | 0.69 | 0.67 |
| | disBERT | **0.75** | 0.73 | **0.76** |
| "Electronics" | MARGOT [28] | 0.53 | 0.74 | 0.61 |
| | MARGOT (TF-IDF) [34] | 0.65 | 0.68 | 0.66 |
| | MARGOT w. BoW | 0.74 | 0.77 | 0.75 |
| | MARGOT w. disc. | **0.77** | 0.71 | 0.74 |
| | $BERT_{base}$ | 0.64 | 0.61 | 0.6 |
| | BERT w. discourse [7] | 0.62 | 0.63 | 0.62 |
| | disBERT | 0.71 | **0.84** | **0.77** |
| "CDs and Vinyl" | MARGOT [28] | 0.54 | 0.77 | 0.64 |
| | MARGOT (TF-IDF) [34] | 0.75 | 0.8 | 0.77 |
| | MARGOT w. BoW | **0.74** | **0.8** | **0.77** |
| | MARGOT w. disc. | 0.76 | 0.7 | 0.73 |
| | $BERT_{base}$ | 0.62 | 0.68 | 0.65 |
| | BERT w. discourse [7] | 0.65 | 0.69 | 0.67 |
| | disBERT | 0.72 | 0.69 | 0.70 |
| AR Dataset (combination of three categories) | disBERT | **0.83** | **0.80** | **0.79** |

Таблица 1 — Экспериментальные результаты качества решения задачи оценки аргументированности текстов для трех категорий набора данных AR и корпуса UKP. Для набора данных AR строка, соответствующая комбинации трех категорий, описывает эксперимент, когда модель обучалась на смешанных текстах из трех категорий.

| Тема | Precision | Recall | $F_1$ score |
|---|---|---|---|
| Abortion | 0.59 | 0.53 | 0.56 |
| Cloning | 0.62 | 0.57 | 0.59 |
| Death penalty | 0.68 | 0.66 | 0.67 |
| Gun control | 0.64 | 0.63 | 0.63 |
| Marijuana legalization | 0.62 | 0.63 | 0.62 |
| Minimum wage | 0.59 | 0.55 | 0.57 |
| Nuclear energy | 0.66 | 0.61 | 0.63 |
| *School uniforms* | *0.67* | *0.55* | *0.60* |

Таблица 2 — Экспериментальные результаты по перекрестной оценке модели для разных тем. Модель была обучена на 7 из 8 тем и протестирована по каждой из восьми тем отдельно. Тема *'School uniforms'* была исключена из обучающего набора.

автоматического выявления аргументационных структур в текста. Однако комбинация модели MARGOT и дополнительных текстовых характеристик превосходит все исследуемые модели на базе BERT. Мы связываем это с тем, что набор данных Debater, который использовался для обучения MARGOT, состоит из текстов публичных дебатов, которые включали в себя аргументированные позиции оппонентов, что близко к распределению аргументированных отзывов пользователей, составляющих набор данных AR Dataset.

Мы также провели сравнение текстов для каждой из трех категорий товаров из AR Dataset. В целом, качество модели практически одинаково для всех трех категорий. Мы можем заметить, что disBERT превосходит все остальные модели по $F_1$-мере для категорий "Movies and TV" и "Electronics". Для категории "CDs and Vinyl" модель-ансамбль, сочетающая функции MARGOT и BoW, немного превосходит disBERT: на 0,02 в точности и на 0,11 в полноте. Точно модель disBERT показывает улучшение по сравнению с $BERT_{base}$. Повышение точности составляет 0,10, а повышение полноты составляет всего 0,01. Модель disBERT повышает производительность модели ансамбля, которая находится на одном уровне с результатами для UKP corpus.

**Пятая глава** диссертации посвящена исследованию более сложного типа задач обработки естественного языка, для решения которых модели также необходимо учитывать дискурсивную структуру текста. К подобному

типу задач можно отнести задачу понимания прочитанного текста (machine reading comprehension, MRC), реферирования и генерации текста, создания вопросно-ответных систем. В данной главе мы концентрируемся на задаче понимания прочитанного текста, для решения которой модели необходимо проанализировать некоторый входной текстовый абзац и ответить на вопрос, заданный по этому абзацу. Для решения задачи MRC в работе предлагается модификация стандартного текстового энкодера, представленного предобученной моделью BERT, посредством добавления в модель дискурсивно-ориентированного механизма внимания. Мы используем граф дискурсивных зависимостей, описанный в предыдущей главе, и добавляем его в дополнительный слой внимания, который будет позволять модели обращать большее внимание на дискурсивно-зависимые части вопроса и входного текста. Экспериментальная оценка показала, что использование дискурс-ориентированного механизма внимания позволяет улучшить качество решения задачи MRC, в особенности, на длинных текстовых абзацах. Модель, описываемая в данной главе, а также результаты проведенного исследования приведены в работе [15].

Модель disBERT, представленная в предыдущей главе, позволяет учитывать дискурсивную структуру текста при решении задачи маскированного языкового моделирования, а также задачи классификации текстов. Однако специфика представления данных, которая требуется для модели disBERT накладывают некоторые ограничения на ее применимость к более сложным языковым задачам.

Модель disBERT работает с тройками, состоящими и пары ЭДЕ и связывающего их дискурсивного отношения, таким образом, для обработки всего текста его изначально необходимо разбить на более мелкие части и рассматривать их независимо друг от друга. Если такой подход применим к задачам классификации, то более сложные задачи, требующие рассмотрения всего текста одновременно могут страдать от подобного представления входных данных. Для решения подобных задач модель должна обрабатывать входную последовательность токенов непрерывно, при этом, анализ дискурсивной структуры может помочь модели концентрироваться на более значимых частях текста при генерации ответа.

В данной главе мы рассматриваем задачу MRC и показываем, что добавления дискурсивной структуры в механизм внимания может повысить ее производительность, обеспечивая более правильное сопоставление между текстовым фрагментом входного абзаца и вопросом, заданного по этому абзацу. В данной главе также исследуется степень влияния дискурсивных, синтаксических и семантических признаков, добавленных в языковую модель индивидуально или в комбинации друг с другом, на качество нахождения ответов на сложные, длинные вопросы, состоящие из нескольких предложений.

На рисунке 4 приведена архитектура предложенной модели BERT, расширенной с помощью дискурсивно-обогащенного механизма внимания.

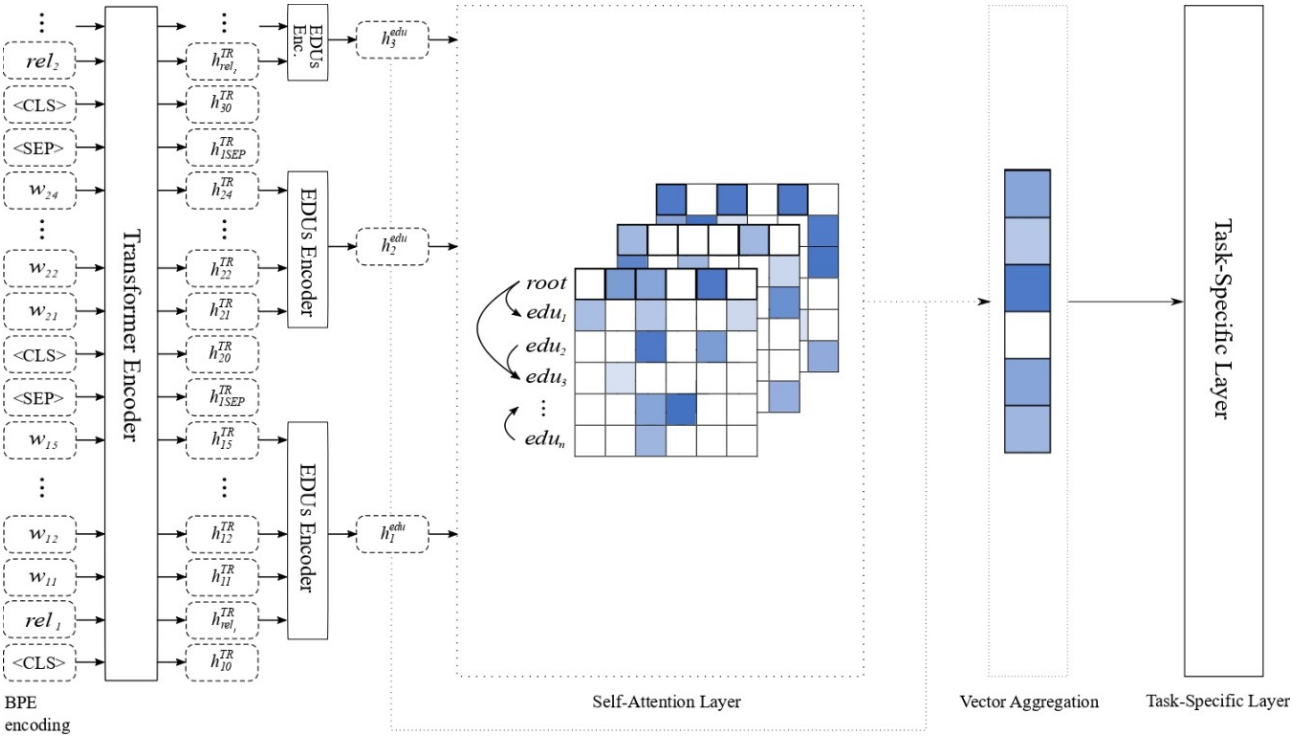


Рисунок 4 — Архитектура модели MRC с дополнительным блоком внимания, учитывающим дискурсивную структуру.

Финальное векторное представление на выходе модели получается простым суммирование контекстуального эмбеддинга входного текста и вопроса ($H$) и дискурсивно-обогащенных эмбеддингов полученных на выходе дополнительного слоя внимания ($H_{disc}$): $\overline{H}_{disc} = H + H'_{disc}$.

В качестве дополнительного исследования мы также рассматриваем степень влияния других лингвистических признаков таких, как синтаксис [58] и семантику [56], на качество решения задачи MRC.

**Экспериментальное исследование.** Мы проверяли разработанную модель с учетом двух наборов данных: первый включает в себя достаточно простые односложные вопросы, для ответа на которые модели не требуется разрешать сложные лингвистические феномены. В качестве примеров подобных вопросов был использован набор данных SQuAD [40, 39]. Вторая часть проведенного экспериментального исследования оценивает способность модели отвечать на более сложные вопросы, для нахождения правильных ответов на которые модели необходимо сначала найти логическую взаимосвязь между частями вопросов, далее установить соответствие между сложным вопросом и входным текстом. В качестве наборов данных с более сложными вопросами были рассмотрены наборы данных NewsQA [48], QA in Context (QuAC) [8] и multi-sentence questions (MSQ) [6]. Результаты для набора данных SQuAD приведены в таблице 3, в таблице 4 приведены результаты для наборов данных со сложными вопросами. В качестве численной оценки приводится значение среднего гармонического между точностью и полнотой полученных ответов, выраженных $F_1$-мерой. Результаты, полученные с помощью предложенной в работе модели, представлены в нижнем блоке таблицы. Результаты лучших из существующих моделей, предложенных для решения задачи MRC на исследуемых наборах данных, приведены в верхнем блоке таблиц; символ * обозначает неопубликованные работы.

**Результаты на наборе SQuAD.** Результаты работы модели на наборе текстов и вопросов SQuAD 1.1 и 2.0 приведена в таблице 3. Базовая модель – это стандартная модель BERT, в которой не учитывается дополнительная лингвистическая информация о тексте. В данной работе мы применяем дообученную модель BERT для кодирования абзаца текста и вопроса к нему, далее полученное векторное представление используется для определения ответа. Кодирование синтаксической, семантической и дискурсивной структуры обеспечивает повышение качества решаемой задачи на 2,2, 3,4 и 3% соответственно. Использование комбинации рассмотренных лингвистических признаков повышает качество нахождения ответов на 5,4%. Отметим, что наименьший прирост качества по сравнению со стандартной моделью BERT достигается при дополнительном кодировании синтаксической структуры текста, что

| Набор данных | v1.1 test | v2.0 test |
|---|---|---|
| | $\mathbf{F}_1$ | $\mathbf{F}_1$ |
| *SQuAD leaderboard* | | |
| FPNet* | - | 93.18 |
| Retro-Reader | - | 92.98 |
| ALBERT | - | 92.20 |
| LUKE* | 95.4 | - |
| Baseline | 88.61 | 83.98 |
| Syntax MRC | 89.90 | 87.13 |
| Semantic MRC | 90.60 | 88.76 |
| **Discourse MRC** | **90.08** | **88.60** |
| **Syntax w. semantic w. discourse MRC** | **93.14** | **90.20** |

Таблица 3 — $F_1$ мера (%) для набора SQuAD 1.1 (v1.1) и SQuAD 2.0 (v2.0). Последняя строка таблицы представляет собой модель, в которой одновременно учитываются все три типа лингвистической информации: синтаксис, семантика и дискурс.

| Набор данных | NewsQA | QuAC | MSQ |
|---|---|---|---|
| | $\mathbf{F}_1$ | $\mathbf{F}_1$ | $\mathbf{F}_1$ |
| *literature + QuAC leaderboard* | | | |
| SpanBERT [25] | 73.6 | - | - |
| DecaProp [46] | 66.3 | - | - |
| RoR* | - | 74.9 | - |
| FlowQA [21] | - | 64.1 | - |
| Baseline | 66.48 | 65.69 | 60.66 |
| Syntax MRC | 70.95 | 71.09 | 66.79 |
| Semantic MRC | 71.84 | 70.15 | 66.55 |
| **Discourse MRC** | **72.13** | **72.40** | **67.80** |
| **Syntax w. semantic w. discourse MRC** | **75.05** | **74.88** | **71.65** |

Таблица 4 — $F_1$ мера (%) для наборов данных со сложными вопросами.

согласуется с результатами синтаксического пробинга моделей, которые показывают, что BERT способен кодировать синтаксис на средних слоях модели. Несмотря на то, что предложенная дискурсивно-обогащенная модель не смогла достичь лучших результатов, по сравнению с другими одиночными и ансамблевыми моделями, предложенными для данной задачи (например, ALBERT и FPNet), данная модель позволила улучшить качество, обеспечиваемое стандартной моделью BERT, не учитывающий лингвистическую структуру.

**Результаты на наборах со сложными вопросами.** В таблице 4 представлены результаты, полученные для более сложных наборов данных NewsQA, QuAC и MSQ. Мы рассматриваем данные наборы совместно, так как вопросы, представленные в них, являются более сложными и детальными, чем вопросы, из набора данных SQuAD. Для того чтобы найти правильный ответ на данный тип вопросов, модели необходимо найти взаимосвязь как между отдельными частями вопроса, так и между вопросом и текстовым абзацем. Отметим, что общее качество решения данных задач падает для всех моделей по сравнению с результатами для набора данных SQuAD, однако прирост качества, связанный с кодированием лингвистической информации, выше, чем для более простых вопросов. Средний вклад синтаксической, семантической и дискурсивной информации составляет 5,3, 5,2 и 6,5% соответственно. Максимальный вклад дискурсивной структуры можно увидеть на наборе данных MSQ. Улучшение качества решения задачи интегрированной системой составляет почти 11% для MSQ и 9,5% в среднем для всех задач. Эти результаты показывают, что чем длиннее и сложнее вопросы, тем выше влияние лингвистической информации, особенно на уровне дискурса. Следует также отметить, что представленная ансамблевая модель превосходит как автономную доработанную модель BERT, так и современные модели для NewsQA и обеспечивает сопоставимые результаты на QuAC.

**Шестая глава** диссертации посвящена вопросам создания интерпретируемых и объяснимых моделей глубинного обучения. В данном исследовании был предложен метод по нахождению текстовых фрагментов, т.н. *текстовых обоснований*, которые способны объяснить, почему модель выдала тот или иной ответ при решении определенной

задачи. Авторы существующих исследований в области разработки интерпретируемых и объяснимых моделей вводят ряд требований, которым должны удовлетворять конструируемые *текстовые обоснования*, а именно: они должны быть *достоверными*, т.е. модель действительно должна полагаться на данные текстовые фрагменты при нахождении решения, более того, они должны быть *интерпретируемыми*, т.е. обоснования должны быть понятны человеку. *Текстовые обоснования*, которые можно получить с использованием предложенного в работе метода удовлетворяют рассмотренным условиям, так как при их построении оцениваются внутренние параметры модели, а именно веса внимания, во-вторых, они являются интерпретируемыми, так как представлены в виде согласованных текстовых фрагментов, расширенных соответствующей им дискурсивной структурой, которая понятна человеку.

После анализа предыдущих работ по интерпретируемости предобученных языковых моделей [22] нами был предложенный улучшенный вариант независимого пайплайна по конструированию *текстовых обоснований* результатов работы языковых моделей. Предложенный пайплайн состоит из двух независимых блоков: блока для извлечения *текстовых обоснований*-кандидатов на основании анализа языковой модели, а также блока предсказания финальных *текстовых обоснований*, конструируемых с помощью анализа дискурсивной структуры текстов и интерпретируемого подхода, основанного на анализе формальных понятий (АФП) [18, 19].

Предлагаемая процедура декомпозиции единой модели интерпретации снижает сложность ее обучения по сравнению с моделями, в которых компонента извлечения и предсказания обучаются совместно [27]. Данный подход совместного обучения требовал оптимизации сложной целевой функции с применением алгоритмов обучения с подкреплением. Предлагаемый метод не требует оптимизации сложной функции, где два компонента непосредственно зависят друг от друга, конструируемые при этом *текстовые обоснования* являются грамматически-согласованными компонентами исходного текста, что обеспечивает высокую степень их объяснимости пользователями данной системы интерпретации. Общая схема работы предложенной модели приведена на рисунке 5.

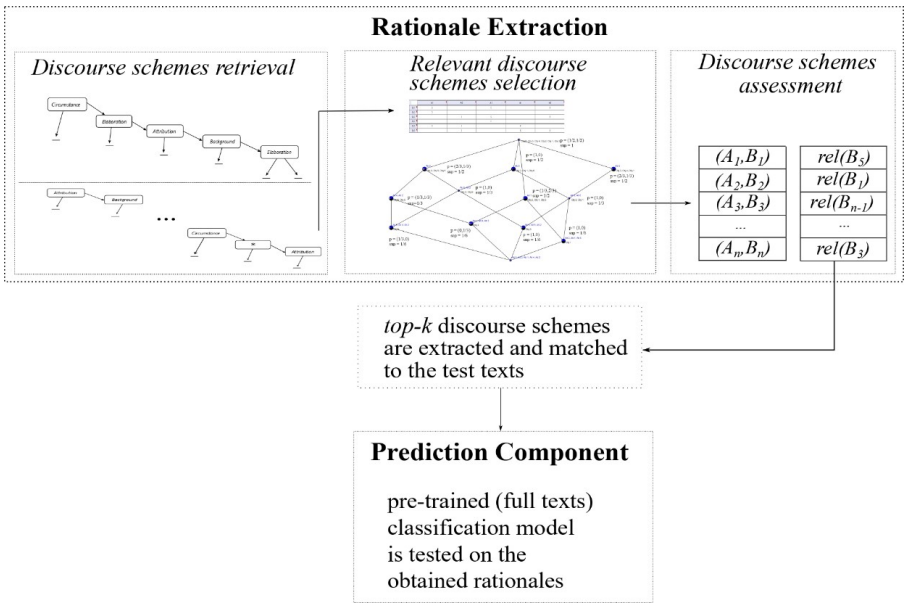


Рисунок 5 — Независимая модель генерации *обоснований* (Independent explanation pipeline, IEP). Блок извлечения обоснования (rationale extraction, RE) используется для нахождения *обоснований*-кандидатов; блок предсказания (rationale prediction, RP) используется для уточнения текстовых обоснований с помощью анализа их дискурсивной структуры.

Формально, *текстовые обоснования* могут быть описаны следующим образом. Пусть *текстовое обоснование* представляет собой определенную комбинацию токенов, извлекаемых из входной последовательности, $rat = \{w_1, w_2, ..., w_K\}$, где $w_k \in W$ и $k \in I$; $W$ – это входная последовательность токенов, а $I$ – это набор индексов, определяющих границы соответствующих фрагментов исходного текста. В идеальном случае, нам необходимо найти такую комбинацию входных токенов, которая бы обеспечила минимальное значение функции потерь для анализируемой задачи. Однако для этого требуется обучить модель на всех возможных комбинациях входных токенов, что неосуществимо с точки зрения вычислительных затрат. Кроме того, требование **интерпретируемости** говорит нам о том, что полученные обоснования должны быть понятны человеку, т.е. они должны быть приведены в виде согласованного текстового фрагмента, а не отдельно стоящих токенов исходного текста.

Разработанный метод нахождения *текстовых обоснований* строится следующим образом. Первый компонент системы отвечает за нахождение значимых частей текста, которым модель приписывает наибольшие веса внимания при генерации ответа. Таким образом, мы можем определить,

какой из текстовых фрагментов оказался более информативным для модели, а каким фрагментом исходного текста можно пренебречь без потери качества.

Далее, запускается компонент предсказания, который служит для извлечения релевантных дискурсивных признаков, которые могут быть использованы для объяснения ответов модели. При нахождении релевантных дискурсивных признаков, мы не учитываем сам текст, но анализируем дискурсивную структуру, соответствующую ему и характеризующуюся графом дискурсивных зависимостей.

Для нахождения дискурсивных признаков, характеризующих текст, для всех текстов обучающей выборки строятся их дискурсивные деревья и соответствующие им графы дискурсивных зависимостей. Далее, из каждого построенного графа дискурсивных зависимостей выделяется так называемый часто-встречаемый подграф $DDG_i$ с помощью алгоритма *gspan* [54]. Данные часто-встречаемые подграфы являются элементарными признаками, описывающими исходный текст. Мы строим бинарные векторы, кодирующие наличие или отсутствие соответствующего часто-встречаемого подграфа, создавая таким образом бинарный контекст $(G,M,I)$, где $G$ – это множество объектов, $M$ – множество признаков (часто-встречаемые подграфы), а $I \subseteq G \times M$ – бинарное отношение, согласно которому если объект $g$ обладает признаком $m$, то $(g,m) \in I$, т.е. подграф $m$ присутствует в дискурсивном графе разбора текста $g$. В соответствии с алгоритмом, предложенным в работе [19], для работы с объектами, описываемыми бинарными признаками, мы строим решетку формальных понятий для положительных и отрицательных примеров в случае бинарной классификации и определяем какие часто-встречаемые дискурсивные подграфы соответствуют положительному, а какие отрицательному классу, таким образом, находя релевантные дискурсивные схемы для решения поставленной задачи.

Задача поиска *текстовых обоснований*, $rat$, превращается в задачу оценки $rel(w_i|DDG_i)$ для каждого $i \in N$, где $rel(w_i|DDG_i)$ – оценка релевантности токена $w_i$, обусловленная его дискурсивной ролью в тексте $DDG_i$. Предполагается, что для задач, зависящих от дискурса, без существенных потерь может быть вычислена оценка $rel(DDG_i)$, а затем может быть выбран $w_i$, соответствующий $rel(DDG_i)$. Мы подтверждаем

это предположение экспериментальным исследованием, проведенным для нескольких задач обработки естественного языка.

Таким образом, весь подход к конструированию *текстовых обоснований* можно описать следующим образом:

1. Определить *текстовые обоснования*-кандидаты как набор значимых с точки зрения механизма внимания текстовых фрагментов.
2. Сгенерировать финальные *текстовые обоснования* с учетом анализа их дискурсивной структуры с применением метода АФП.
3. Провести предсказание ответа моделью, принимающей на вход только *текстовые обоснования*.

**Экспериментальное исследование.** Экспериментальное исследование работы предложенного подхода к извлечению *текстовых обоснований* проводилось для трех классификационных задач обработки естественного языка: оценки аргументированности текстов (AR и UKP Corpus), анализа тональности текстов (Stanford Sentiment Treebank (SST) [44], Movies) и многоклассовой классификации текстов по темам (AgNews [10]).

В наборе данных SST документы разделены на два класса в зависимости от их тональности: положительные и отрицательные. В наборе доступно 9613 документов. Набор данных Movies – это набор данных отзывов пользователей, также размеченных в зависимости от их тональности. В данном наборе присутствует 1999 документов. Набор данных AgNews представляет собой набор данных классификации документов на несколько классов, в котором новостные статьи помечены одной из соответствующих тем: *Спорт*, *Наука*, *Бизнес* и *Мировые новости*. Этот набор данных содержит 127600 документов.

Стандартная модель $\text{BERT}_{base}$, моделт *disBERT* и BERT с дискурсивным механизмом внимания используются в качестве базовых для извлечения *обоснований*. Мы предполагаем, что модели, ориентированные на анализ дискурсивной структуры, должны показывать лучшие результаты для извлеченных *обоснований*, поскольку они были обучены учитывать дискурсивную структуру во время обучения.

В таблице 5 представлены результаты, полученные рассмотренными моделями для *обоснований*, построенными с использованием различных методов, предложенных в работах [27, 2, 22]. Мы также приводим результаты, полученные моделями при обработке несокращенных текстов (**Full text**), чтобы оценить достоверность построенных *обоснований*.

| Модель | RE | AR | UKP | SST | Movies | AGNews |
|---|---|---|---|---|---|---|
| $\text{BERT}_{base}$ | Full text | 0.60 | 0.35 | 0.90 | 0.94 | 0.96 |
| | Lei et al. | 0.52 | 0.33 | 0.74 | **0.92** | 0.87 |
| | Bastings et al. | 0.51 | 0.28 | 0.59 | 0.72 | – |
| | Att.-based | 0.63 | 0.32 | **0.81** | 0.91 | **0.94** |
| | IEP | **0.64** | **0.34** | 0.71 | 0.80 | 0.82 |
| disBERT | Full text | 0.68 | 0.54 | 0.67 | 0.74 | 0.76 |
| | Lei et al. | 0.52 | 0.45 | 0.54 | 0.62 | 0.67 |
| | Bastings et al. | 0.51 | 0.48 | 0.60 | **0.65** | —- |
| | Att.-based | 0.63 | 0.52 | **0.61** | 0.59 | **0.68** |
| | IEP | **0.65** | **0.53** | 0.53 | 0.57 | 0.62 |
| BERT ext. with discourse-aware SAN | Full text | 0.77 | 0.69 | 0.89 | 0.87 | 0.85 |
| | Lei et al. | 0.62 | **0.65** | 0.74 | **0.92** | **0.87** |
| | Bastings et al. | 0.69 | 0.57 | 0.59 | 0.52 | -— |
| | Attn-based | 0.53 | 0.52 | 0.71 | 0.71 | 0.82 |
| | IEP | **0.72** | 0.63 | **0.75** | 0.67 | 0.81 |

Таблица 5 — Результаты работы моделей по извлечению *текстовых обоснований* для разных наборов данных.В таблице приведена численное значение $F_1$ меры.

Проанализировав результаты, приведенные в таблице, можно сделать вывод о том, насколько хорошо сгенерированные *обоснования* помогают модели принимать решение (чем выше значения $F_1$ меры, тем информативнее найденные *обоснования*), а также проанализировать для каких задач найденные с помощью анализа дискурса обоснования окажутся наиболее информативными.

Результаты работы модели *disBERT* и модели BERT с дискурсивно-ориентированным механизмом внимания для текстов по задаче оценки аргументированности, построенных только по найденным *обоснованиям*, почти сравнимы с результатами этих моделей при генерации предсказаний на полных текстах. Однако, можно отметить, что для наборов данных для задачи анализа тональности, использование сокращенных текстов, состоящих только из сгенерированных *обоснований* снизило качество решения задачи. Например, на наборе Movies $F_1$ мера упала на 0,17, что говорит о том, что предложенный подход к извлечению обоснований больше подходит для более сложных дискурсивно-зависимых

задач обработки естественного языка. Мы также можем заметить, что предложенный подход к извлечению *обоснований* достигает лучшего показателя $F_1$ меры в комбинации с моделью disBERT, учитывающей дискурсивную структуру текста в процессе обучения. В целом, можно отметить, что качество работы моделей, протестированных на сокращенных текстах, состоящих только из извлеченных *обоснований*, незначительно ниже, чем у полных текстов (например, 0,68 против 0,65 для модели disBERT и предложенного подхода по извлечению *обоснований*). Это указывает на то, что построенные обоснования достоверны и кодируют большую долю значимой информации из исходного текста. Комбинация предложенного подхода к извлечению *обоснований* и стандартной модели $\text{BERT}_{base}$ улучшает результаты, полученные моделью $\text{BERT}_{base}$ на полных текстах, на наборе данных AR, что свидетельствует о том, что компонент извлечения *обоснований* смог не только выделить информативные текстовые фрагменты, но и удалить нерелевантные/зашумленные части входного текста.

**Заключение** резюмирует основные аспекты диссертации, излагая выводы, полученные в рамках проведенного исследования.

## ВЫВОДЫ

К основным выводам, полученным при проведении данного исследования, можно отнести следующее:

1. Был проведен анализ способности предобученных языковы моделей на базе архитектуры Трансформер кодировать дискурсивную структуру текста, в рамках которого выявлено, что не все предобученные модели одинаково хорошо учитывают данную лингвистическую структуру.
2. Были предложены и реализованы две модификации для существующей предобученной модели BERT (модель disBERT и BERT с дискурсивно-ориентированным механизмом внимания).
3. Была проведена экспериментальная оценка предложенных моделей, показывающая их применимость к задачам обработки естественного языка, таких как анализ аргументированности текста и понимание прочитанного текста. Для всех задач модели,

учитывающие дискурс, превзошли результаты стандартной модели BERT.

4. Был предложен подход к генерации *текстовых обоснований*, способных объяснить результаты работы предобученных языковых моделей.
5. Предложенные в работе модели и размеченные наборы данных выложены в открытый доступ в репозитории GitHub.

# Список использованных источников


[1] Yossi Adi et al. "Fine-grained analysis of sentence embeddings using auxiliary prediction tasks". In: *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*. 2017.

[2] Joost Bastings, Wilker Aziz, and Ivan Titov. "Interpretable neural predictions with differentiable binary variables". In: *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. 2020. DOI: 10.18653/v1/p19-1284.

[3] Yonatan Belinkov et al. "Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 1–10. URL: https://aclanthology.org/I17-1001.

[4] Yonatan Belinkov et al. "What do neural machine translation models learn about morphology?" In: *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*. Vol. 1. 2017. DOI: 10.18653/v1/P17-1080.

[5] Terra Blevins, Omer Levy, and Luke Zettlemoyer. "Deep RNNs encode soft hierarchical syntax". In: *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*. Vol. 2. 2018. DOI: 10.18653/v1/p18-2003.

[6] Laurie Burchell et al. "Querent Intent in Multi-Sentence Questions". In: arXiv:2010.08980 (2020). arXiv: 2010.08980 [cs.CL].

[7] Tuhin Chakrabarty et al. "AmperSand: Argument mining for persuasive online discussions". In: *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. 2019. DOI: 10.18653/v1/d19-1291.

[8] Eunsol Choi et al. "QUAC: Question answering in context". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*. 2020. DOI: 10.18653/v1/d18-1241.

[9] Kevin Clark et al. "What Does BERT Look at? An Analysis of BERT's Attention". In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 276–286. DOI: 10.18653/v1/W19-4828. URL: https://aclanthology.org/W19-4828.

[10] Gianna M. Del Corso, Antonio Gullí, and Francesco Romani. "Ranking a stream of news". In: *14th International World Wide Web Conference, WWW2005*. 2005. DOI: 10.1145/1060745.1060764.

[11] Daniel Cristea, Nancy Ide, and Laurent Romary. "Veins Theory : A Model of Global Discourse Cohesion and Coherence". In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and of the 18th International Conference on Computational Linguistics COLING98ACL98* 1 (1998).

[12] Jacob Devlin et al. "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*. Vol. 1. 2019.

[13] Boris Galitsky. "Learning Discourse-Level Structures for Question Answering". In: *Developing Enterprise Chatbots* (2019), pp. 177–219. DOI: 10.1007/978-3-030-04299-8_7.

[14] Boris Galitsky, Dmitry Ilvovsky, and Elizaveta Goncharova. "On a Chatbot Conducting Dialogue-in-Dialogue". In: *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. Stockholm, Sweden: Association

for Computational Linguistics, Sept. 2019, pp. 118–121. DOI: 10.18653/v1/W19-5916. URL: https://aclanthology.org/W19-5916.

[15] Boris Galitsky, Dmitry Ilvovsky, and Elizaveta Goncharova. "Relying on Discourse Analysis to Answer Complex Questions by Neural Machine Reading Comprehension". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. Held Online: INCOMA Ltd., Sept. 2021, pp. 444–453. URL: https://aclanthology.org/2021.ranlp-1.51.

[16] Boris Galitsky, Dmitry Ilvovsky, and Elizaveta Goncharova. "Relying on Discourse Trees to Extract Medical Ontologies from Text". In: *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer-Verlag, 2021, pp. 215–231. ISBN: 978-3-030-86854-3. DOI: 10.1007/978-3-030-86855-0_15. URL: https://doi.org/10.1007/978-3-030-86855-0_15.

[17] Boris A. Galitsky, Sergei O. Kuznetsov, and Daniel Usikov. "Parse thicket representation for multi-sentence search". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 7735 LNCS. 2013. DOI: 10.1007/978-3-642-35786-2_12.

[18] Bernhard Ganter and Rudolf Wille. "Formal concept analysis". In: Springer Verlag, Berlin, 1998.

[19] Elizaveta Goncharova and Sergei Kuznetsov. "Increasing the Efficiency of Packet Classifiers with Closed Descriptions". In: *In proceedings of the 7th International Workshop "What can FCA do for Artificial Intelligence"? co-located with International Joint Conference on Artificial Intelligence, FCA4AI@IJCAI*. 2019, pp. 75–86.

[20] André Grüning and Andrej A. Kibrik. "Modelling referential choice in discourse: a cognitive calculative approach and a neural network approach". In: *Anaphora processing: linguistic, cognitive and computational modelling* (2005).

[21] Hsin Yuan Huang, Wen Tau Yih, and Eunsol Choi. "FlowQA: Grasping flow in history for conversational machine comprehension". In: *7th International Conference on Learning Representations, ICLR 2019*. 2019.

[22] Sarthak Jain et al. “Learning to Faithfully Rationalize by Construction”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4459–4473. DOI: 10.18653/v1/2020.acl-main.409. URL: https://aclanthology.org/2020.acl-main.409.

[23] Ekaterina Jasinskaja, Jörg Mayer, and David Schlangen. “Discourse Structure and Information Structure: Interfaces and Prosodic Realization”. In: *Interdisciplinary Studies on Information Structure (ISIS)* (2004).

[24] Yangfeng Ji and Jacob Eisenstein. “Representation learning for text-level discourse parsing”. In: *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*. Vol. 1. 2014. DOI: 10.3115/v1/p14-1002.

[25] Mandar Joshi et al. “SpanBERT: Improving Pre-training by Representing and Predicting Spans”. In: *Transactions of the Association for Computational Linguistics* 8 (2020). DOI: 10.1162/tacl_a_00300.

[26] Fajri Koto, Jey Han Lau, and Timothy Baldwin. “Discourse Probing of Pretrained Language Models”. In: 2021. DOI: 10.18653/v1/2021.naacl-main.301.

[27] Tao Lei, Regina Barzilay, and Tommi Jaakkola. “Rationalizing Neural Predictions”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 107–117. DOI: 10.18653/v1/D16-1011. URL: https://aclanthology.org/D16-1011.

[28] Marco Lippi and Paolo Torroni. “MARGOT: A web server for argumentation mining”. In: *Expert Systems with Applications* 65 (2016). DOI: 10.1016/j.eswa.2016.08.050.

[29] Annie Louis, Aravind Joshi, and Ani Nenkova. “Discourse Indicators for Content Selection in Summaization”. In: Sept. 2010, pp. 147–156.

[30] T.P. Makhalova et al. “FCA-based approach for interactive query refinement with IR-chatbots”. In: *In proceedings of the Russian Advances in Artificial Intelligence, RAAI 2020*. Vol. 2648. 2020, pp. 144–156.

[31] William Mann and Sandra Thompson. "Rhetorical Structure Theory: Toward a functional theory of text organization". In: *Text* 8 (3 1988). DOI: 10.1515/text.1.1988.8.3.243.

[32] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*. 2013.

[33] Pramod K. Mudrakarta et al. "Did the model understand the question?" In: *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*. Vol. 1. 2018. DOI: 10.18653/v1/p18-1176.

[34] Marco Passon et al. "Predicting the Usefulness of Amazon Reviews Using Off-The-Shelf Argumentation Mining". In: *Proceedings of the 5th Workshop on Argument Mining*. 2018, pp. 35–39. DOI: 10.18653/v1/w18-5205.

[35] Matthew E. Peters et al. "Deep contextualized word representations". In: *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*. Vol. 1. 2018. DOI: 10.18653/v1/n18-1202.

[36] Rashmi Prasad et al. "The Penn Discourse TreeBank 2.0." In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA), May 2008. URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf.

[37] Alec Radford et al. "Improving Language Understanding by Generative Pre-Training". In: *Preprint* (2018).

[38] Colin Raffel et al. "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *Journal of Machine Learning Research* 21 (2020).

[39] Pranav Rajpurkar, Robin Jia, and Percy Liang. "Know what you don't know: Unanswerable questions for SQuAD". In: *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*. Vol. 2. 2018. DOI: 10.18653/v1/p18-2124.

[40] Pranav Rajpurkar et al. "SQuAD: 100,000+ questions for machine comprehension of text". In: *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*. 2016. DOI: 10.18653/v1/d16-1264.

[41] Nils Reimers et al. "Classification and clustering of arguments with contextualized word embeddings". In: *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. 2019. DOI: 10.18653/v1/p19-1054.

[42] Corby Rosset et al. "Knowledge-Aware Language Model Pretraining". In: *ArXiv* abs/2007.00655 (2020).

[43] Xing Shi, Inkit Padhi, and Kevin Knight. "Does string-based neural MT learn source syntax?" In: *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*. 2016. DOI: 10.18653/v1/d16-1159.

[44] Richard Socher et al. "Recursive deep models for semantic compositionality over a sentiment treebank". In: *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. 2013.

[45] Christian Stab et al. "Cross-topic argument mining from heterogeneous sources". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*. 2018. DOI: 10.18653/v1/d18-1402.

[46] Yi Tay et al. "Densely connected attention propagation for reading comprehension". In: *Advances in Neural Information Processing Systems*. Vol. 2018-December. 2018.

[47] Ian Tenney et al. "What do you learn from context? Probing for sentence structure in contextualized word representations". In: *7th International Conference on Learning Representations, ICLR 2019*. 2019.

[48] Adam Trischler et al. "NewsQA: A Machine Comprehension Dataset". In: 2017. DOI: 10.18653/v1/w17-2623.

[49] Ashish Vaswani et al. "Attention is all you need". In: *Advances in Neural Information Processing Systems*. Vol. 2017-December. 2017.

[50] Alex Wang et al. “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding”. In: *EMNLP 2018 - 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Proceedings of the 1st Workshop*. 2018. DOI: 10.18653/v1/w18-5446.

[51] Alex Wang et al. “SuperGLUE: A stickier benchmark for general-purpose language understanding systems”. In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019.

[52] Florian Wolf and Edward Gibson. “Representing Discourse Coherence: A Corpus-Based Study”. In: *Computational Linguistics* 32 (June 2005), pp. 249–287.

[53] Jiacheng Xu et al. “Discourse-Aware Neural Extractive Text Summarization”. In: 2020. DOI: 10.18653/v1/2020.acl-main.451.

[54] Xifeng Yan and Jiawei Han. “gSpan: Graph-based substructure pattern mining”. In: *Proceedings - IEEE International Conference on Data Mining, ICDM*. 2002. DOI: 10.1109/icdm.2002.1184038.

[55] Zhuosheng Zhang et al. “Semantics-Aware BERT for Language Understanding”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 2020. DOI: 10.1609/aaai.v34i05.6510.

[56] Zhuosheng Zhang et al. “Semantics-Aware BERT for Language Understanding”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (05 2020). DOI: 10.1609/aaai.v34i05.6510.

[57] Zhuosheng Zhang et al. “SG-Net: Syntax-Guided Machine Reading Comprehension”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (05 2020). DOI: 10.1609/aaai.v34i05.6511.

[58] Zhuosheng Zhang et al. “SG-Net: Syntax-Guided Machine Reading Comprehension”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (05 2020). DOI: 10.1609/aaai.v34i05.6511.

[59] Wei Zhao et al. “Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance”. In: *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. 2019. DOI: 10.18653/v1/d19-1053.

[60] Yukun Zhu et al. “Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books”. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015. DOI: 10.1109/ICCV.2015.11.