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

Факультет гуманитарных наук

Программа подготовки бакалавров по направлению по направлению 45.03.03. Фундаментальная и прикладная лингвистика

Татаринов Максим Дмитриевич

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Оценка применимости метода детектирования семантических изменений слов нейросетевой языковой моделью на основе генерируемых определений

Руководитель
канд. комп. н.
А. В. Демидовский

Научный консультант
канд. фил. н., доцент департамента фундаментальной и прикладной лингвистики
А. Ю. Хоменко

Нижний Новгород, 2024

Оглавление

Введение

На **актуальность** настоящей работы указывают следующие факторы. Во-первых, активное изучение темы автоматического определения семантических изменений. В последние годы в работах использовались различные методы, включая статические эмбеддинги, контекстуальные эмбеддинги и заканчивая генерацией определений с помощью языковых моделей в новейших исследованиях [1—3] . При этом, абсолютное большинство исследований, посвященных моделированию определений, проводятся с использованием материала английского языка [4]. Для русского языка вопрос анализа семантических изменений на основе автоматически сгенерированных определений недостаточно изучен. Во-вторых, неудовлетворительное качество традиционных методов для основных потенциальных пользователей таких технологий, таких как лексикографы, историки языка и социологов. Например, лексикографам недостаточно данных только о факте сдвига значения, им хотелось бы получать описания старых и новых значений слов в пригодной для чтения форме, возможно, даже с дополнительными пояснениями. Данная проблема может решаться моделированием определений с использованием языковых моделей, при использовании которых исследователи смогут получить более наглядные результаты [3].

Целью настоящей работы является оценка применимости метода детектирования семантических изменений слов нейросетевой языковой моделью на основе генерируемых определений.

Из поставленной цели были сформулированы следующие **задачи**:

1. Провести анализ существующей литературы и решений по задаче детектирования семантических изменений на основе генерируемых определений.

2. Собрать датасет словарей русского языка в качестве материала для обучения модели.
3. Обучить языковую модель на данных словарей для того, чтобы генерировать определения.
4. Создать алгоритм автоматического определения семантических сдвигов на основе векторного представления.
5. Провести анализ метрик и качества обученной языковой модели и сравнить их с существующими решениями.
6. Создать алгоритм визуализации результатов.
7. Провести качественный анализ результатов работы компьютерной программы.

Объектом исследования является метод детектирования семантических изменений слов.

Предметом исследования является применимость метода детектирования семантических изменений слов с использованием нейросетевой языковой модели на основе генерируемых определений.

Для решения поставленных задач были использованы следующие **методы**:

1. Метод анализа и синтеза для создания теоретической базы для данного исследования на основе литературы.
2. Компьютерный метод для написания алгоритмов программы и обучения модели.
3. Методы обработки естественного языка для предобработки текстов.
4. Методы глубокого обучения для алгоритма автоматического определения семантических сдвигов на основе их векторного представления.
5. Метод лексико-семантического анализа (используется при оценке визуализаций алгоритма).

Новизна настоящей работы состоит в том, что для детектирования семантических изменений значений слов применяется на материале русского языка и с использованием SOTA-моделей.

Практическая значимость данной работы заключается в том, что результаты настоящей работы можно применять для определения степени семантического сдвига лексем, с наличием визуализаций и определений для каждого выявленного значения, что может быть использовано в лексикологии, где необходимы актуальные данные для построения новых словарей [3]. Кроме того, модель, позволяющая автоматически генерировать качественные словарные определения, может быть полезна в таких задачах обработки естественного языка, как анализ тональности, машинный перевод и разграничение семантической неоднозначности [4].

В качестве **материала исследования** используется диахронический корпус НКРЯ, охватывающий три периода (1700—1916, 1918—1991 и 1992—2016 годы) и имеющий в совокупности 250 миллионов словоупотреблений. Данный корпус выбран, поскольку датасет слов для валидации с изменившимся и неизменившимся значением, использующийся для оценки алгоритма, основан на данном корпусе [5]. Корпус был получен по запросу к авторам НКРЯ.

Апробация работы. Основные положения настоящей работы были представлены на конференции (VIII Всероссийская научная студенческая конференция НИУ ВШЭ – Нижний Новгород «Цифровые технологии в современной молодежной науке», 17 апреля 2024 г., тема доклада: «Оценка применимости метода детектирования семантических изменений слов нейросетевой языковой моделью на основе генерируемых определений»).

Глава 1. Теоретические аспекты автоматического выявления семантических изменений

1.1. Понятия и классификации

В рамках изучения исторических изменений в лексике языка или языков, лингвисты оперируют такими понятиями, как лексические изменения, семантические изменения, грамматикализация и лексическая замена [6].

Лексические изменения в широком смысле охватывают все виды диахронических преобразований в словарном составе языка, в то время как в более узком значении термин относится к появлению новых форм в языке, таких как заимствованные слова и неологизмы, а также к устаревшим словам.

Семантические изменения или семантический сдвиг являются особым случаем лексических изменений, когда существующая форма (лексема) приобретает или теряет конкретное значение, что приводит к увеличению или уменьшению полисемии.

Примером таких изменений может служить эволюция английских слов, когда ранее специализированное слово для обозначения определенного вида собаки стало общим термином (dog), в то время как более раннее общее слово для «собаки» — современный аналог которого hound — сейчас используется для обозначения специального вида собак.

Грамматикализация описывает особый вид семантических изменений, когда слова с полным значением превращаются в служебные слова и, в конечном итоге, в связанные грамматические морфемы. Примером может служить развитие глагольного аффикса *-ся* из безударного возвратного местоимения формы винительного падежа.

В рамках настоящей работы мы будем заниматься исследованием семантических изменений лексического значения слов, не касаясь иных схожих явлений.

Чтобы подробно разобрать изменения семантики слова, полезно обратиться к типологии, разработанной американским лингвистом Леонардом Блумфилдом [7, 8]. Он выделил следующие типы:

1. Сужение значения (narrowing), при котором слово начинает употребляться в более узких сферах общения за счет конкретизации своего значения. Например, Old English mete, означавшее «еда», сузилось до современного английского meat, означающего «мясо».
2. Расширение значения (widening), при котором значение слова становится более общим, способным применяться во многих сферах общения. Например, слово dog в Middle English означало «собаку определенной породы», но теперь описывает «любую собаку».
3. Гипербола (hyperbole), при которой значение слова нарочно преувеличивается. Например, слово из допрефранцузского *ex-tonare, означавшее «ударить громом», преобразовалось в французское étonner, означающее «удивлять».
4. Литота (litotes), когда значение слова нарочно подвергается преуменьшению. Например, Old English cwellan, которое означало «убить», произошло от предшествующего *['kwalljan], означавшего «пытать».
5. Деградация значения (degeneration), когда значение слова приобретает более негативное значение в течение времени. Например, Old English cnafa, первоначально обозначало «мальчик, слуга», и превратилось в knave (лжец).
6. Возвышение значения (elevation), при котором значение слова приобретает более позитивное значение в течение времени. Например, knight произошло от Old English cniht, которое означало «мальчик, слуга».

7. Метафора (metaphor), заключающаяся в изменении значения на основе скрытого сравнения. Например, прагерманское *['bitraz], означавшее «колючий», превратилось в bitter, означающее «горький».
8. Перенос значения, или метонимия (metonymy), на основе смежности или близости ассоциативных связей. Например, Old French joue, означавшее «щека», стало означать «челюсть».
9. Синекдоха (synecdoche), когда значение слова представляет отношение части к целому или наоборот. Например, прагерманское *['tu:naz] означало «забор» и превратилось в английском в «небольшой город».

1.2. Обзор существующих методов

1.2.1. Предыстория

Традиционно для изучения изменений семантики слов использовались ручные методы детального анализа текстов. Из существующих исследований истории значений слов в русском языке можно привести исследование 1500 слов и 5000 связанных с ними выражений В.В. Виноградова [9], а также книгу «Два века в двадцати словах», в деталях описывающую историю набора двадцати слов. [10] Хотя ручные методы продолжают применяться в лингвистике, в последнее время появилось множество полуавтоматических и автоматических методов, способных расширить и углубить такие исследования, а также упростить их проведение. Одним из факторов, позволивший предлагать такие решения стала цифровизация документов в различных областях и появление диахронических корпусов. [6].

1.2.2. Статические эмбеддинги

До 2020 года в работах чаще используют статические эмбеддинги.

Для русского языка в качестве примера можно привести проект Shiftry [11], в котором для анализа семантических сдвигов использовались модели Word2Vec, выровненные методом Прокруста. Эти модели были обу-

чены на обширном корпусе русскоязычных новостных текстов, охватывающем период с 2010 по 2020 годы, и позволили отследить диахронические изменения в употреблении слов. Для подсчета степени семантического сдвига использовалось косинусное расстояние между векторами из различных временных срезов.

Статические эмбеддинги оставались наиболее актуальными вплоть до 2020, где показывали лучшие результаты в SemEval-2020 Task 1 [12]. Они эффективно моделируют значение слов в зависимости от обучающего корпуса без опоры на объемные предобученные модели, превосходя по этому качеству модели, основанные на совстречаемости слов. Среди недостатков можно отметить необходимость большого объема слов в корпусах для стабильности эмбеддингов; необходимость выравнивания моделей, обученных на отдельных наборах данных, соответствующим временным срезам, что может вносить шум; кроме того, они моделируют среднее значение слова на основе его употребления в корпусе, не позволяя различать разные значения слова.

1.2.3. Контекстуализированные эмбеддинги

Статические модели вложений слов присваивают каждому слову (лемме) один и тот же вектор независимо от контекста, в то время как современные достижения в области обработки естественного языка позволили разработать модели, обеспечивающие получение контекстуализированных представлений высокого качества. Данные модели отличаются тем, что на этапе вывода токенам присваиваются различные вложения в зависимости от их контекста, что позволяет проводить более глубокий диахронический анализ языковых изменений с использованием контекстуализированных векторных представлений слов.

Применение контекстуализированных векторных представлений задало новый стандарт для высококачественных, чувствительных к контексту

представлений в обработке естественного языка. В статье, где исследователи использовали предварительно обученные модели BERT и ELMo, настроенные на полном корпусе Русского национального корпуса, было обнаружено, что эти модели показывают значительную корреляцию с человеческими оценками при определении диахронического семантического изменения слов в русском языке [2]. Использовались алгоритмы, такие как косинусное сходство по прототипам слов и методы кластеризации, для выявления семантических сдвигов.

Одни из последних работ по теме автоматического выявления семантических сдвигов для русского языка были написаны в рамках соревнования RuShiftEval, прошедшего в 2021 году [5]. В ходе него участники должны были рассмотреть три исторических периода русского языка и общества: предсоветский (1700-1916), советский (1918-1990) и постсоветский (1992-2016). Исследование базировалось на наборе данных RuShiftEval, который состоит из 111 русских существительных (99 в тестовом наборе и 12 в наборе для разработки), вручную аннотированных по степени изменения их значения в трех парах временных периодов.

Аннотаторам предлагалось оценить семантическую связь значений целевого слова в двух предложениях из разных временных периодов. Оценки (от 1 до 4) отражали степень семантического родства между значениями слова, где 1 обозначало отсутствие связи между значениями, а 4 – их совпадение. Затем индивидуальные оценки усреднялись, формируя общую меру семантической родственности между употреблениями слова в разные временные периоды. Такая задача как правило называется Word-in-Context или WiC.

Для каждого из 99 целевых русских слов участники должны были представить три значения, соответствующих семантическому изменению в упомянутых парах временных периодов. Эти значения использовались для построения трех ранжирований: RuSemShift1, RuSemShift2 и RuSemShift3.

В качестве метрики оценки использовалась ранговая корреляция Спирмена между ранжированием слов, сгенерированным системой, и золотым ранжированием, полученным в ручной аннотации.

Победители вышеупомянутого соревнования (команда GlossReader) указывают, что проблемой в существующих решениях являлось то, что эмбеддинги несут в основном информацию о форме слова, а не значении [13]. Чтобы решить это, они дообучали модель XLM-R на задаче генерации эмбеддингов, максимально близким к таким, какие получены на соответствующим использованиям слов словарным определениям [14].

При дообучении их система включает в себя два отдельных энкодера на основе XLM-R: Энкодер контектов для кодирования предложения с целевым словом и энкодер глоссов для кодирования определения слова. Система оценивает возможные значения смысла слова путём сравнения векторных представлений слова и его определений. При этом для обучения использовались данные только по английскому языку, но модель также показала хорошие результаты для русского языка.

Далее, исследователи получали эмбеддинги контекстов слов с помощью дообученного энкодера контекстов, высчитывали расстояние с помощью различных метрик расстояния, самым эффективным из которых были евклидово расстояние с нормализацией, после чего логистическая регрессия приводила значения к формату в датасете.

Авторы статьи предоставляют доступ к части исходного кода их исследования [15].

Так, были опубликованы следующие компоненты:

1. Код, предназначенный для генерации прогнозов на основе заранее вычисленных эмбеддингов, полученных с использованием модели.
2. Код для оценки результатов.

В то же время, авторы исследования не представили в открытый доступ следующие части:

1. Код для предварительного обучения модели.
2. Код, позволяющий осуществлять инференцию для получения контекстуализированных эмбеддингов, сформированных на основе предварительно обученной модели.

В соответствии с инструкциями, данными авторами, мы запустили доступный код, в следствие чего были полуены высокие результаты, совпадающие с тем, что сообщают авторы в своей работе:

Таблица 1.1.. Коэффициэнты корреляции

Пары периодов	Коэффициент корреляции
Среднее	0.8021
pre-Soviet:Soviet	0.7808
Soviet:post-Soviet	0.8032
pre-Soviet:post-Soviet	0.8223

Среди недостатков работы можно отметить неспособность модели корректно выявлять значения тех слов, которые отличаются от ближайших аналогов в английском, например, «пионер», связанный с коммунистической идеологией и не соответствующий в полной мере слову «scout».

Кроме того, команда DeepMistake представила решение, занявшее в соревнование второе место [16]. Однако, они смогли доработать его и повысить результаты до первого уже после окончания соревнования.

Исследователи обучали модель XLM-R на обширном многоязычном датасете Word-in-Context, а затем дооубчали ее на наборе данных RuSemShift для настоящей задачи, приводит к наилучшим результатам. В отношении архитектуры авторы утверждают, что применение линейного слоя на верхнем уровне, основанного на объединении L1-метрики и скалярного произведения между контекстуализированными эмбеддингами XLM-R, показывает лучшую производительность по сравнению с более традицион-

ными подходами, такими как конкатенация эмбеддингов и использование нелинейных классификаторов.

Исследователи выложили исходный код полностью и предлагают возможность воспроизвести их результат [17]. Значени метрик, сообщенные исследователями, воспроизводятся.

Таблица 1.2.. Коэффициэнты корреляции с использованием IsoReg

Пары периодов	Коэффициент корреляции
Среднее	0.8494
pre-Soviet:Soviet	0.8563
Soviet:post-Soviet	0.841
pre-Soviet:post-Soviet	0.8511

Среди недостатков статьи можно выделить то, что авторы не предоставляют возможность визуализации или интерпретации результатов, кроме непосредственно получившегося значения метрики.

Тем не менее, применимость таких методов была подвергнута сомнению в работе Giulianelli et al. [3], где ставится под сомнение широкая практичность ранее упомянутых подходов. Они утверждают, что такие методы практически неинтерпретируемы, поскольку они не дают описаний значений слов, а лишь бинарные результаты наличия или отсутствия семантического изменения. Исследование, которое в наибольшей степени занимается этой проблемой, - это GlossReader, где исследователи предлагают способ визуализации и интерпретации результатов. Однако у этого метода есть свои недостатки, обсуждаемые выше. Учитывая эти факты, новый подход, включающий моделирование определений, вызывает интерес для задачи обнаружения семантических изменений.

Начало интереса к моделированию определений как теме исследования в области обработки естественного языка можно отнести к работе

Noraset et al. [18]. Они были среди первых, кто исследовал потенциал использования векторных представлений слов для автоматической генерации определений. Изначально была поставлена упрощенная задача с моносемантическими словами, которые, как правило, имеют одно значение и, следовательно, одно определение. Однако оставалась нерешенной проблема многозначных слов. Gadetsky et al. выделили важное условие для моделирования определений: необходимость контекста для точного захвата нюансов языка [19]. В своем исследовании они включили примеры предложений для предоставления контекста модели, что оказалось решающим шагом в возможности модели справляться с полисемией и улучшении ее производительности.

Несмотря на достижения, сделанные вышеупомянутыми исследователями, область моделирования определений все еще сталкивалась с значительными проблемами. Huang et al. выявили наличие таких проблем, как проблема слов вне словаря, когда модели сталкиваются с трудностями в работе со словами, не встречавшимися во время обучения, а также проблемы избыточной и недостаточной специфичности в определениях [20]. Исследователи сообщают: «Избыточно специфичные определения представляют узкие значения слов, в то время как недостаточно специфичные определения представляют общие и нечувствительные к контексту значения.» Huang et al. решили эти проблемы, используя предварительно обученную модель энкодера-декодера, а именно Text-to-Text Transfer Transformer (T5), и ввели механизм ранжирования, предназначенный для тонкой настройки специфичности генерируемых определений. Метод был протестирован на стандартных наборах данных для оценки и показал значительное улучшение по сравнению с предыдущими методами.

Самой актуальной работой по теме использования сгенерированных большими языковыми моделями определений для автоматического выявления семантических изменений является статья Giulianelli et al. [3].

Авторы определяют задачу генерации определений следующим образом: для заданного слова w и примера использования s (предложения, содержащего w) необходимо сгенерировать определение d на естественном языке, которое будет грамматически корректным и точно передавать значение слова w в контексте его использования. Для генерации определений они используют модель Flan-T5, версию трансформера T5, дополнительно обученную на 1,8 тысячах задач по обработке естественного языка. Ниже вы можете видеть пример работы модели.

Таблица 1.3.. Пример определения, сгенерированного моделью Flan-T5 XL

Пример использования	‘Примерно половина солдат в наших стрелковых взводах были призывниками, которых мы обучали около шести недель.’
Целевое слово	призывник
Сгенерированное определение	‘ЧЕЛОВЕК, КОТОРЫЙ ПОДЛЕЖИТ ПРИЗЫВУ В ВООРУЖЕННЫЕ СИЛЫ’

Первым шагом исследователи выбирают, используя метрики BLEU, NIST, BERTScore, наиболее подходящий под задачу промпт из нескольких вариантов, например ”what is the definition of <trg>?” или ”define the word <trg>”.

Для дообучения модели авторы используют три датасета, каждый из которых содержит определения слов, сопровождаемые примерами употребления: WordNet, данные Оксфордского словаря и CoDWoE, основанный на определениях и примерах, извлеченных из Викисловаря.

Для оценки качества модели исследователи используют метрики SacreBLEU, ROUGE-L и BERT-F1.

Для демонстрации работы со сгенерированными определениями авторы работы используют датасет, в котором слова представлены в графах диахронного использования слов (Diachronic Word Usage Graphs, DWUG), взвешенных, ненаправленных графах, узлами которых служат примеры использования слов, а веса рёбер отражают семантическую близость пар употреблений. DWUG созданы на основе многоэтапного процесса человеческой аннотации, в ходе которого аннотаторы оценивали семантическую связность пар употреблений слов по 4-балльной шкале.

Прежде всего, авторы исследования проводят анализ корреляции между близостью пар слов в DWUG и контекстуальными эмбеддингами токенов, эмбеддингами предложений примеров использования, а также сгенерированными определениями. Результаты показали, что сгенерированные определения обладают более высокой степенью корреляции с данными из DWUG, чем традиционно полученные эмбеддинги.

Далее исследователи анализируют пространство эмбеддингов определений слов, чтобы выяснить, как они могут помочь в различении разных значений слов. Они обнаружили, что эмбеддинги определений образуют более плотные и четко определенные кластеры по сравнению с эмбеддингами токенов и примеров предложений, что делает их подходящими для представления значений слов.

Далее авторы присваивали кластерам, полученным на основе данных из DWUG, соответствующие им определения. Для обобщения определений в одном кластере авторы использовали самое прототипическое из них. Они представляли все определения с помощью их эмбеддингов предложений и выбирали в качестве прототипичного определение, эмбеддинг которого наиболее похож на среднее значение всех эмбеддингов в кластере.

Авторы приходят к выводу, что сгенерированные определения слов могут играть роль семантического представления слов, аналогичному традиционным эмбеддингам. Они находят большие языковые модели достаточно

развитыми для генерации определений простым промптом. При этом полученные таким образом определения превосходят по качеству традиционные эмбеддинги и являются более наглядными.

Глава 2. Имплементация автоматического выявления семантических изменений

2.1. Обучение языковой модели на данных тезауруса

В качестве модели была выбрана FRED-T5-1.7B, являющаяся одной из новейших языковых моделей, выпущенных SberDevices и обученных с нуля на материале русского языка [21]. Для выбора модели мы использовали бенчмарк для оценки продвинутого понимания русского языка ”RussianSuperGLUE” [22].В бенчмарке присутствуют шесть групп задач, охватывая общую диагностику языковых моделей и различные лингвистические задачи: понимание здравого смысла, логическое следование в естественном языке, рассуждения, машинное чтение и знания о мире. FRED-T5-1.7B занимает самое высокое место в лидерборде данного бенчмарка, со значением 0.762, уступая лишь результатам выполнения данных заданий людьми со значением 0.811, что свидетельствует о ее способности к выдающемуся языковому пониманию и анализу. Таким образом, FRED-T5-1.7B представляется наиболее подходящей языковой моделью для задачи генерации определений.

Одной из ключевых особенностей модели FRED-T5-1.7B является наличие денойзеров. Денойзеры — это специальные механизмы, задача которых состоит в очистке текста от шума, то есть в восстановлении удаляемых или искажаемых частей текста. В модели используется семь различных денойзеров, каждый из которых выполняет уникальную функцию в процессе обучения. Основные задачи денойзеров включают в себя восстановление удаленных участков текста, а также продолжение текстовых последовательностей.

В настоящей работе при работе с моделью используется денойзер, помеченный спецтокеном ”<LM>”, который задействован в задаче продолжения последовательности текста.

Действия, описанные далее, подкрепляются кодом, выложенным в открытый доступ, на сайте GitHub и могут быть воспроизведены. [23]

В качестве материала, используемого для обучения модели, выступил датасет словарей, который включал в себя материал из русской версию Викисловаря [24], а также из МАС «Малого академического словаря» [25]. Материал Викисловаря получен с помощью написанного скрипта на языке Python, позволяющего извлечь данные из выгрузки Викисловаря в формат JSONL. Материал Малого академического словаря был получен с помощью скраппера, написанного на языке Python. В загруженном наборе данных в каждом вхождении присутствовали идентификатор статьи, лексема, про которую написана данная статья, а также определения с примерами использования.

Таблица 2.1.. Информация о лексеме из Викисловаря

Лексема	**Определения и примеры использования**
прозябнуть	**сильно озябнуть, промёрзнуть:** Я и без того прозяб, инстинкт тянет меня согреться, а какой-то нелепый долг повелевает лезть в холодную воду. Усталость возьмет свое, тогда можно жестоко прозябнуть и опасно заболеть. **прорасти:** Сперва надо его в землю посадить, потом ожидать, покуда в нем произойдет процесс разложения, потом оно даст росток, который прозябнет, в трубку пойдет, восколосится и т. д.

Полученный материал был очищен от вхождений, не имеющих при себе примеров использования, информативных определений, например, «Состояние по знач. глаг. линять», или не содержащих определений вовсе, а также имеющие такие определения, которые представляют грамматическую информацию о слове вместо лексического значения, например, «наречие к причастию приглашающий». На выходе было получено 270 тысяч 555 вхождений.

Примеры и слова были отформатированы под формат запроса модели. В начале после слова «Контекст» шел пример использования слова, после чего шла фраза «Определение слова», в которую включалось само слово. Таким образом, на вход модель принимает лексему и контекст, в которой она употреблялась, а на выход ожидается сгенерированное определение.

Таблица 2.2.. Пример отформатированного запроса модели

Поле	**Значение**
input_text	<LM>Контекст: ”Усталость возьмет свое, тогда можно жестоко прозябнуть и опасно заболеть.” Определение слова ”прозябнуть”:
target_text	Сильно озябнуть, промёрзнуть.

FRED-T5-1.7B была дообучена на полученном из Викисловаря материале в течение трёх эпох с постоянным шагом обучения 0.001, размером батча 16 и оптимизатором Adafactor на одной видеокарте RTX 3090. Для ускорения обучения и экономии видеопамяти использовалась технология LoRa со следующими параметрами: r – 32, alpha – 64, dropout – 0.1, что позволило уменьшить количество обучаемых параметров до 14155776 (0.8% от общего числа параметров), что поволило сэкономить используемую видеопамять и ускорить обучение. Более подробный обзор гиперпараметров модели, а также хода ее обучения доступен в приложении А.

Для оценки качества обучения модели используются метрики BLEU и ROUGE-L, которые оценивают формальную схожесть текста: BLEU оценивает точность совпадений n-грамм в сгенерированном тексте по сравнению с эталонным текстом [26], а ROUGE-L измеряет схожесть между сгенерированным текстом и эталонным текстом на основе наибольшей общей последовательности слов [27]. Также использовалась метрика BERT-F1, учитывающая семантику сравниваемых текстов благодаря использованию контекстуальных эмбеддингов при подсчете значения метрики [28]. Использование нескольких метрик позволяет получить более полную картину качества модели, поскольку каждая из них оценивает разные аспекты сгенерированного текста. Как традиционные BLEU и ROUGE-L, так и более современный BERT-F1 активно используются в задачах обработки естественного языка, в том числе в задачах генерации текста. В данной работе использовались версии этих инструментов, взятые из библиотеки evaluate [29]. Так, в обзорной статье по моделированию определений утверждается, что на момент выпуска статьи BLUE использовался в 9 научных публикациях, ROUGE-L и BERTScore – в 3 [4]. Кроме того, данные метрики используются и в более новых работах. Так, в настоящей статье результаты данных метрик будут сравниваться с таковыми из статьи Giulianelli M. et al., где сообщаются результаты трёх вышеперечисленных метрик при обучении модели T5 для задаче генерации определений на английском языке [3].

Таблица 2.3.. Результаты дообучения FRED-T5-1.7B на датасете Викисловаря

Метрика	Значение
BLEU	8.3162
ROUGE-L	37.45
BERT-F1	78.06

2.2. Тестирование модели на материале соревнования Rushifteval

С помощью модели были получены определения для тестовой части датасета соревнования Rushifteval.

Для векторизации сгенерированных определений использовалась paraphrase-multilingual-mpnet-base-v2, векторы были нормализированы, после чего расстояние между векторным представлением определений считалось с помощью косинусного расстояния. Результат приводился в формат значений датасета с помощью линейной регрессии, тренированной на датасете Rusemeval.

Таблица 2.4.. Коэффициэнты корреляции с использованием IsoReg

Пары периодов	Коэффициент корреляции
Среднее	0.72
pre-Soviet:Soviet	0.707
Soviet:post-Soviet	0.731
pre-Soviet:post-Soviet	0.723

Рассмотрим данные результаты в сравнении с аналогами из соревнования Rushifteval.

Таблица 2.5.. Результаты алгоритма в сравнении с результатами команд Rushifteval.

Команда	**досоветский: советский**	**советский: постсовет-ский**	**досоветский: постсовет-ский**	**Среднее**
GlossReader	0.781	0.803	0.822	0.802
DeepMistake	0.798	0.773	0.803	0.791
vanyatko	0.678	0.746	0.737	0.720
FRED-T5-FN	**0.707**	**0.731**	**0.723**	**0.72**
aryzhova	0.469	0.450	0.453	0.457
Discovery	0.455	0.410	0.494	0.453
UWB	0.362	0.354	0.533	0.417
dschlechtweg	0.419	0.373	0.383	0.392
jenskaiser	0.430	0.310	0.406	0.382
SBX-HY	0.388	0.281	0.439	0.369
Baseline	0.314	0.302	0.381	0.332
svart	0.163	0.223	0.401	0.262
BykovDmitrii	0.274	0.202	0.307	0.261
fdzr	0.217	0.251	0.065	0.178

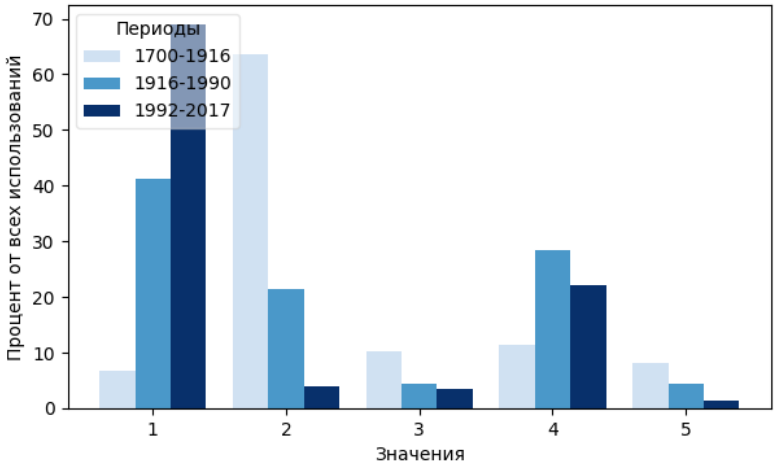
Как видно из таблицы, настоящее решение лучше по качеству большинства аналогов из соревнования Rushifteval, хоть и уступает некоторым, использующим модели XLM-R. Два решения с самым высоким качеством были описаны ранее в Главе 1.

2.3. Визуализация результатов работы модели

Для создания визуализаций семантических изменений слов используются используются библиотеки Matplotlib и Scikit-learn. Полученные с помощью модели определения векторизуются по аналогии с выше описанными главами. Так как для слов, имеющих одинаковое значение, модель склонна генерировать семантичсеки близкие, однако не идентичные дословно определения, для группировки таких схожих определений применяется алгоритм кластеризации DBSCAN из библиотеки Scikit-learn на основе векторных представлений. Алгоритм кластеризации может настраиваться вручную через два ключевых параметра: ”eps» и «min_samples”. Параметр «eps» определяет максимальное расстояние между двумя точками, чтобы они считались находящимися в одном соседстве. ”Min_samples» определяет минимальное количество точек, которые должны образовывать плотно связанную группу, чтобы она образовывала кластер. Затруднительно сказать заранее, какие параметры кластеризации подойдут для визуализации каждого конкретного слова. Представляется хорошим вариантом сначала выбирать небольшие значения и после повышать их, пока значения, сформулированные похожим, но разным образом, не объединяться в единые кластеры. После этого, для каждого полученного кластера выбирается прототипическое определение, векторное представление которого наиболее близко к центру кластера. Данное определение выбирается для описания данного значения (кластера). Затем библиотека Matplotlib применяется для создания столбиковых диаграмм, отражающих частоту употреблений различных значений слова во времени, и для обеспечения наглядности с помощью цветовой градации и легенд, содержащих прототипические определения каждого из значений.

Результатом анализа является график по типу иллюстрации, где представлена столбчатая диаграмма, показывающая процентное соотношение

значений исследуемого слова за разные периоды времени. Каждая категория обозначена на диаграмме своим цветом и соответствующим временным интервалом: светло-синий цвет для 1700-1916, средне-синий для 1916-1990 и темно-синий для 1992-2017. Под диаграммой находится расшифровка значений, а также использованные параметры визуализации.



1: Место, куда свозят, сваливают мусор, отходы.
2: Столкновение, драка.
3: Беспорядочное скопление кого-либо, чего-либо.
4: Место, где свалены, нагромождены какие-л. предметы.
5: Беспорядочное, беспорядочное скопление людей.

Параметры: eps=0.9, min_samples=30

Рис. 2.1.. Изменение значений слова "Свалка"

Глава 3. Анализ результатов работы модели

Для дальнейшего анализа результатов алгоритма использовались 20 слов с изменившемся значением из книги «Два века в двадцати словах» [10]. Использования данных слов брались из диахронического корпуса НКРЯ.

Из каждого периода (досоветсткий, советский и постсоветсткий) бралась выборка из 300 вхождений, где для каждого использования слова генерировалось определение, а после строился график по аналогии с описанием визуализации выше.

Далее для каждого слова описана семантика слов на основе словарей в соответствии с рекомендациями издания И.А. Стернина [30]. В качестве материала будут взяты «Большой толковый словарь» (далее *БТС*) [31], «Толковый словарь русского языка Дмитриева» (далее *ТСРЯ*) [32] и книга «Два века в двадцати словах». После чего будет проведено сравнение выявленных при семантическом описании лексемы значений и тех, что выявлены алгоритмом.

Кроме того, произведено сравнение статистической информации по использованию слов в разные периоды для значений, соотносимых со значениями из книги «Два века в двадцати словах».

Следует учитывать то, что в книге исследуются периоды длиной меньше, чем в настоящей работе. Например, вместо досоветского выделяют 1800-1849, 1850-1874, 1875-1899, а также 1900-1924, в связи с чем не представляется возможным выявить изменения между короткими периодами из книги.

Рассмотрим 20 слов внимательнее.

Свалка

Семантическое описание лексемы *Свалка* по словарям

В результате анализа семем лексемы *свалка* в толковых словарях были выделены семь групп значений, которые можно условно сформулировать следующим образом:

1. Место для сбора мусора, нечистот. (*«Место, куда свозят, выбрасывают мусор, нечистоты, негодные вещи.»* в БТС, *«Место, куда вывозят, выбрасывают мусор, нечистоты, негодные вещи.»* в ТСРЯ, *«Место для сбора мусора, нечистот»* в «Два века в двадцати словах»)
2. Процесс сваливания. (*«к Свалить»* в БТС, *«Процесс сваливания»* в «Два века в двадцати словах»)
3. Всеобщая драка. (*«Свалкой называют всеобщую драку, в которой участвует много людей»* в ТСРЯ, *«Драка»* в «Два века в двадцати словах»)
4. Скопление людей, толпа. (*«Скопление людей, толпа.»* в «Два века в двадцати словах» и в БТС)
5. Груда, куча, нагромождение чего-либо. (*«Беспорядочно накиданная груда, куча чего-л.»* и *«Если кто-либо превращает квартиру в свалку, то это означает, что там в беспорядке нагромождаются предметы, мебель и пр.»* в БТС, *«Свалкой называют беспорядочно накиданную груду каких-либо предметов.»* в ТСРЯ, *«Груда»* в «Два века в двадцати словах»)
6. Битва. (*«Битва»* в «Два века в двадцати словах»)

Результат алгоритма

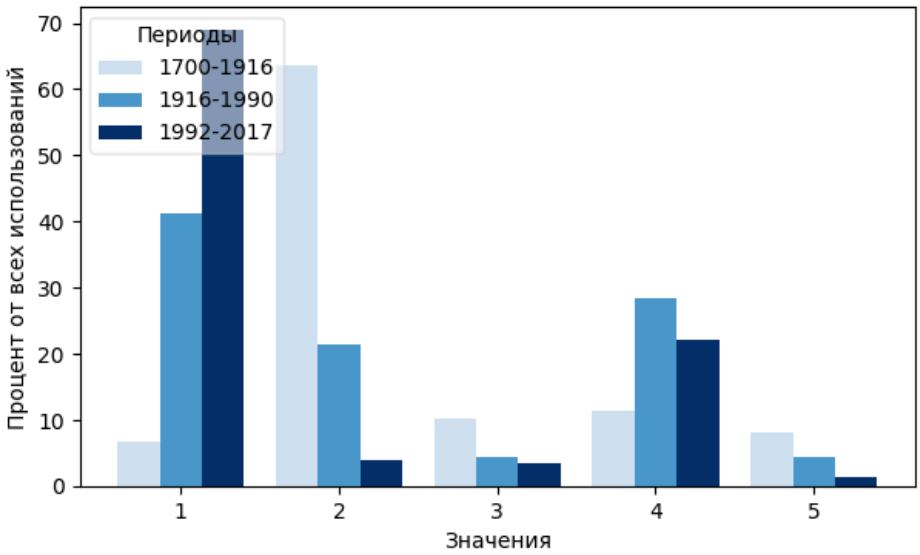


Рис. 3.1.. Изменение значений слова *Свалка* (Параметры: eps=0.9, min_samples=50)

Значения для визуализации слова *Свалка*:

1. Место, куда свозят, сваливают мусор, отходы.
2. Столкновение, драка.
3. Беспорядочное скопление кого-либо, чего-либо.
4. Место, где свалены, нагромождены какие-л. предметы.
5. Беспорядочное, беспорядочное скопление людей.

Анализ результатов

Первые четыре определения корректно сформулированы. Пятое имеет повторение слова «беспорядочное» – ошибку в генерации определения. Далее мы будем считать это определение как «Беспорядочное скопление людей.».

Три определения имеют явные аналоги в составленном ранее описании значений слова. ’Место, куда свозят, сваливают мусор, отходы.’ имеет с ’Место, куда свозят, выбрасывают мусор, нечистоты, негодные

вещи.’ общие смысловые элементы: «место», «своз/сваливание», «отходы/нечистоты».

’Столкновение, драка.’ соответствует определению ’Всеобщая драка, потасовка.’, хоть в словаре оно и более узкое из-за наличия помимо смыслового элемента «драка» семы «всеобщности».

’Беспорядочное скопление людей.’ имеет аналог ’Скопление людей, толпа.’. Здесь общие семы «скопление», «люди», хоть и имеется дифференциальная сема «беспорядочности», делающая определение алгоритма более узким.

’Место, где свалены, нагромождены какие-л. предметы.’ близко к ’Беспорядочно накиданная груда, куча чего-л.’ Общими семами являются «нагромождение», «предметы», однако не акцентируется «беспорядочность» явно, вместо этого эта сема присутствует в значениях слова «нагромождены» («Нагромоздить – построить в чрезмерно большом количестве, очень тесно или в беспорядке.»).

Также алгоритм предлагает более общее значение ’Беспорядочное скопление кого-либо, чего-либо.’, которое может включать как живые, так и неживые объекты, объединяя значения ’Скопление людей, толпа.’ и ’Груда, куча, нагромождение чего-либо.’

Алгоритм не предлагает значение ’Процесс сваливания’. В этом случае акцентируется действие «свалить», указывающее на процесс перемещения предметов/материала с целью создания свалки. Однако, это значение не вынесено отдельно ни в одном периоде в «Двух веках в двадцати словах» и является редким, что могло быть причиной отсутствия в визуализации.

Кроме того, в визуализации отсутствует значение ’Битва’, которое указывается как преобладающее для периода до 1850 года. В предсказаниях модели присутвуют примеры с этим значением, однако их количество в исследуемом материале незначительно, поэтому оно не вошло в визуализацию. Например, для *«Но в свалке, как обыкновенно действует кавалерия,*

сабля или палаш лучше» было сгенерировано определение ’Бой, в котором участвуют несколько противников’.

Перейдем к частотности значений.

В книге «Два века в двадцати значениях» как появившееся в 1900-ых годах указано значение *«Место для сбора мусора, помойка.»*, соответствующее первому значению, предложенному алгоритмом ’Место, куда свозят, сваливают мусор, отходы.’. Как видно из графика результатов алгоритма, оно почти не используется в досоветский период, но становится главным с 40% использования в советский период и доминирует в постсоветский с около 70%. Эти данные совпадают с тем, что говорится в книге, где утверждается 87% использования значения *«Помойка.»* в 1998-1997 годы, 32% для 1925-1949 годов.

Уменьшается же судя по графику преимущественно значение 2 (’Столкновение, драка.’), которое падает с 65% использований в досоветский период до 5% в постсоветский. В книге резульаты схожи. Так, утверждается, что в 1875-1899 году слово имело значение ’Драка.’ в 71% использований, а к 1998-1997 значение упало до 12%.

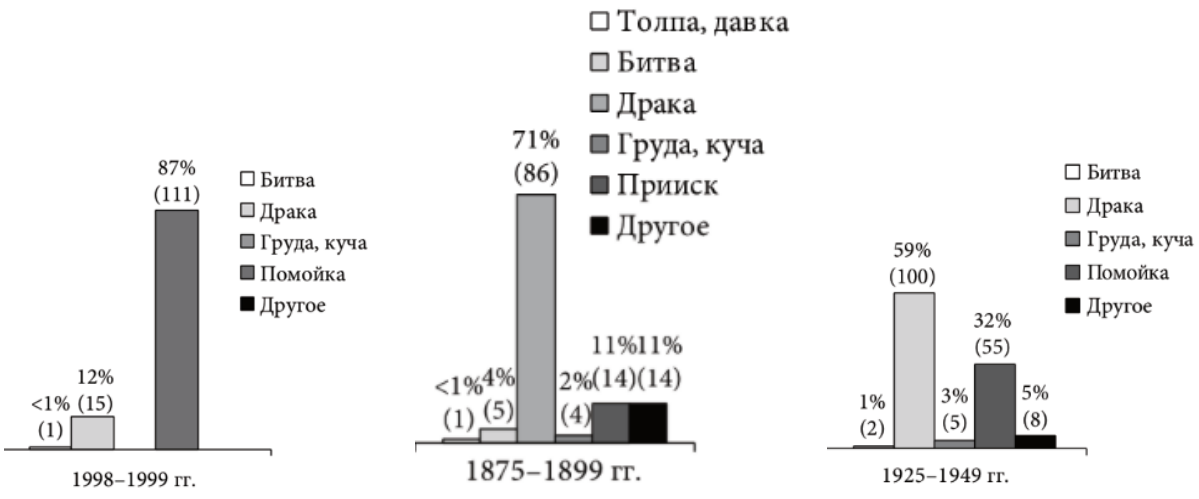


Рис. 3.2.. Визуализации для слова "Свалка" из книги "Два века в двадцати словах".

Таким образом, модель довольно точно отражает реальное изменение значений слова «свалка» во времени, согласуясь с данными из толкового словаря и историческим исследованием. Она адекватно выделяет как наиболее широко используемое сегодня значение, связанное с местом сбора мусора, так и менее очевидные значения, включая драку и беспорядочное скопление предметов или людей.

Заключение

XXX

Результаты настоящей работы можно применять для определения степени семантического сдвига лексем, с наличием визуализации и определений для каждого выявленного значения, что может быть использовано в лексикологии, где необходимы актуальные данные для построения новых словарей [3]. Кроме того, модель, позволяющая автоматически генерировать качественные словарные определения, может быть полезна в таких задачах обработки естественного языка, как анализ тональности, машинный перевод и разграничение семантической неоднозначности [4]. Кроме того, возможно обучение модели большего размера, что повысит качество генерации.

Ограничениями подхода можно считать необходимость в значительных вычислительных ресурсах. Несмотря на то, что FRED-T5-1.7B запускается на ЦПУ, запуск на большом количестве вхождений займет значительное число времени. Для запуска на ГПУ же необходима видеокарта с 8 ГБ видеопамяти.

Код, использованный во время выполнения настоящей работы, выложен в открытый доступ на сайте GitHub и может быть воспроизведен. [23]

Список литературы

1. *Kutuzov A.*, *Øvrelid L.*, *Szymanski T.*, *Velldal E.* Diachronic word embeddings and semantic shifts: a survey // Proceedings of the 27th International Conference on Computational Linguistics. — Santa Fe, New Mexico, USA : Association for Computational Linguistics, 08.2018. — С. 1384—1397.
2. *Rodina J.*, *Trofimova Y.*, *Kutuzov A.*, *Artemova E.* ELMo and BERT in semantic change detection for Russian. — 2020.
3. *Giulianelli M.*, *Luden I.*, *Fernández R.*, *Kutuzov A.* Interpretable Word Sense Representations via Definition Generation: The Case of Semantic Change Analysis. — 2023.
4. *Gardner N.*, *Khan H.*, *Hung C.-C.* Definition modeling: literature review and dataset analysis // Applied Computing and Intelligence. — 2022. — Т. 2. — С. 83—98.
5. *Pivovarova L.*, *Kutuzov A.* RuShiftEval: a shared task on semantic shift detection for Russian //. — 06.2021. — С. 533—545.
6. Computational approaches to semantic change. — Berlin : Language Science Press, 2021.
7. *Bloomfield L.* Language. — New York : Holt, Rinehart, Winston, 1933.
8. *Harris T. M.* Semantic Shift in the English Language //. — 2014.
9. *Виноградов В.*, *Шведова Н.* История слов: около 1500 слов и выражений и более 5000 слов, с ними связанных. — Институт русского языка им. В.В. Виноградова РАН, 1999.

10. *Данова М. К.*, *Добрушина Н. Р.*, *Опачанова А. С.* [и др.]. Два века в двадцати словах. — Москва : Издательский дом Высшей школы экономики, 2018. — 455 с. ; — Электронное издание. Системные требования: Adobe Reader XI либо Adobe Digital Editions 4.5; экран 10".

11. *Kutuzov A.*, *Fomin V.*, *Mikhailov V.*, *Rodina J.* SHIFTRY: WEB SERVICE FOR DIACHRONIC ANALYSIS OF RUSSIAN NEWS //. — 01.2020. — С. 500—516.

12. *Schlechtweg D.*, *McGillivray B.*, *Hengchen S.*, *Dubossarsky H.*, *Tahmasebi N.* SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection // CoRR. — 2020. — Т. abs/2007.11464.

13. *Rachinskiy M.*, *Arefyev N.* Zero-shot Cross-lingual Transfer of a Gloss Language Model for Semantic Change Detection //. — 06.2021. — С. 578—586.

14. *Conneau A.*, *Khandelwal K.*, *Goyal N.*, *Chaudhary V.*, *Wenzek G.*, *Guzmán F.*, *Grave E.*, *Ott M.*, *Zettlemoyer L.*, *Stoyanov V.* Unsupervised Cross-lingual Representation Learning at Scale // CoRR. — 2019. — Т. abs/1911.02116.

15. GlossReader. — URL: https://github.com/myrachins/RuShiftEval (дата обр. 18.01.2024).

16. *Arefyev N.*, *Fedoseev M.*, *Protasov V.*, *Panchenko A.*, *Homskiy D.*, *Davletov A.* DeepMistake: Which Senses are Hard to Distinguish for a Word-in-Context Model //. — 06.2021. — С. 16—30.

17. DeepMistake. — URL: https://github.com/Daniil153/DeepMistake (дата обр. 18.01.2024).

18. *Noraset T.*, *Liang C.*, *Birnbaum L.*, *Downey D.* Definition Modeling: Learning to define word embeddings in natural language. — 2016.

19. *Gadetsky A.*, *Yakubovskiy I.*, *Vetrov D.* Conditional Generators of Words Definitions // 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. 266—271.

20. *Huang H.*, *Kajiwara T.*, *Arase Y.* Definition Modelling for Appropriate Specificity // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. — Online, Punta Cana, Dominican Republic : Association for Computational Linguistics, 11.2021. — С. 2499—2509.

21. *Zmitrovich D.* [и др.]. A Family of Pretrained Transformer Language Models for Russian. — 2023.

22. *Shavrina T.*, *Fenogenova A.*, *Anton E.*, *Shevelev D.*, *Artemova E.*, *Malykh V.*, *Mikhailov V.*, *Tikhonova M.*, *Chertok A.*, *Evlampiev A.* RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). — Association for Computational Linguistics, 2020.

23. *Tatarinov M. D.* Work Definition Modeling. — 2024 ; — Accessed: 2024-04-24. https://github.com/tatarinovst2/work-definition-modeling.

24. *Wiktionary contributors*. Wiktionary, the free dictionary. — 2023. — URL: https://www.wiktionary.org/ (дата обр. 10.04.2024).

25. Словарь русского языка: В 4-х т. — Москва : Русский язык, 1981-1984. — В 4-х томах.

26. *Papineni K.*, *Roukos S.*, *Ward T.*, *Zhu W. J.* BLEU: a Method for Automatic Evaluation of Machine Translation. — 2002. — Окт.

27. *Lin C.-Y.* ROUGE: A Package for Automatic Evaluation of summaries //. — 01.2004. — С. 10.

28. *Zhang T.*, *Kishore V.*, *Wu F.*, *Weinberger K. Q.*, *Artzi Y.* BERTScore: Evaluating Text Generation with BERT. — 2020.

29. Evaluate. — URL: https://github.com/huggingface/evaluate (дата обр. 15.11.2023).

30. *Стернин И. А.*, *Рудакова А. В.* Словарные дефиниции и семантический анализ. — Воронеж, 2017. — С. 34.

31. *Кузнецов С. А.* Большой толковый словарь русского языка: А-Я. — СПб. : Норинт, 1998. — С. 1534. — РАН. Ин-т лингв. исслед. Сост., гл. ред. канд. филол. наук С. А. Кузнецов.

32. *Ахапкин Д.* [и др.]. Толковый словарь русского языка : Ок. 2000 слов-вар. ст., свыше 12000 значений. — Москва : Астрель [и др.], 2003. — С. 989. — ГУП ИПК Ульян. Дом печати.

Приложение А. Обучение модели

Таблица А.1.. LoRa параметры

Параметр	Значение
r	32
lora_alpha	64
lora_dropout	0.1

Таблица А.2.. Trainer параметры

Параметр	Значение
learning_rate	1e-3
lr_scheduler_type	constant
batch_size	16
gradient_checkpointing	true
gradient_accumulation_steps	1
weight_decay	0.01
optimizer	adafactor
num_train_epochs	4

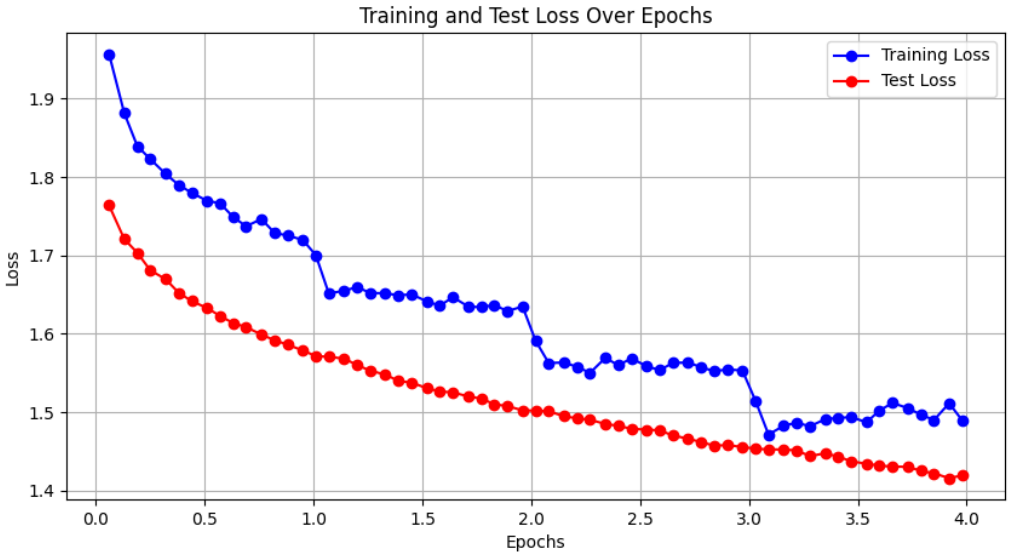


Рис. А.1.. Лосс при обучении модели