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

Детектирование семантических изменений слов нейросетевой языковой моделью на основе генерируемых определений

Рецензент
к. филол. н., доц.
М.А. Климова

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

Традиционно для изучения изменений семантики слов использовались ручные методы детального анализа текстов (Виноградов, Шведова, 1999), (Данова [et al.], 2018). Такие исследования включали в себя многолетнее изучение источников и труд множества исследователей. Однако, сегодня цифровизация документов и появление текстовых баз данных открыли новые перспективы для исследований, облегчая доступ к текстам и позволяя разработать полуавтоматические и автоматические методы анализа (Tahmasebi [et al.], 2021). Настоящая работа посвящена одному из таких перспективных методов, который возможно использовать при анализе семантических изменений, – моделированию определений.

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 (Kutuzov, Øvrelid [et al.], 2018), (Rodina [et al.], 2020), (Giulianelli [et al.], 2023). При этом, абсолютное большинство исследований, посвященных моделированию определений, проводятся с использованием материала английского языка (Gardner [et al.], 2022). Для русского языка вопрос анализа семантических изменений на основе автоматически сгенерированных определений недостаточно изучен. Во-вторых, неудовлетворительная интерпретируемость традиционных методов для основных потенциальных пользователей таких технологий, таких как лексикографы, историки языка и социологи. Например, лексикографам недостаточно данных только о факте сдвига значения, им предпочтительно 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 (Giulianelli [et al.], 2023).

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

Для достижения поставленной цели были сформулированы следующие **задачи**:

1. провести анализ существующей литературы и решений по задаче детектирования семантических изменений на основе генерируемых определений,
2. собрать обучающий датасет на основе словарей русского языка,
3. обучить языковую модель для того, чтобы генерировать определения,
4. разработать 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. разработать алгоритм визуализации результатов,
7. провести качественный анализ результатов разработанного алгоритма.

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. методы программирования 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.

Новизна настоящей работы состоит в том, что детектирование семантических изменений значений слов применяется на материале русского языка с использованием метода моделирования определений и с использованием современных 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 (Giulianelli [et al.], 2023). Кроме того, модель, позволяющая автоматически генерировать качественные словарные определения, может быть полезна в таких задачах обработки естественного языка, как анализ тональности, машинный перевод, извлечение лексических смыслов и разграничение семантической неоднозначности (Gardner [et al.], 2022) Например, извлечение лексических смыслов используется в Sketch engine, однако данный алгоритм неспособен давать описания значений, вместо этого сообщая слова из того же семантического поля(URL: https://www.sketchengine.eu/guide/word-sense-induction/).

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. Корпус был получен по запросу к авторам НКРЯ.

В первой главе исследования рассмотрены теоретические аспекты автоматического выявления семантических изменений, включая обзор существующих на данный момент методов решения задачи автоматического детектирования семантических изменений. Во второй главе дано общее описание предлагаемого подхода. Третья глава включает в себя описание особенностей и деталей реализации предлагаемого подхода, вычислительные эксперименты, а также качественный анализ полученных результатов.

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. Понятие семантических изменений

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

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 устареванию форм в языке, а также появлению новых, таких как заимствованные слова и неологизмы (Tahmasebi [et al.], 2021).

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 (Tahmasebi [et al.], 2021).

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.

Лексическая замена представляет собой явление, когда одно слово или выражение вытесняется другим, часто синонимичным, в языке. Например, в английском языке слово *happy* изначально означало ’быть удачливым’, но затем стало означать ’счастливый’. Обратный процесс описывается на примере слова *gay*, которое раньше означало ’счастливый’, а затем стало использоваться исключительно для обозначения гомосексуальности. Этот процесс можно рассматривать как лексическую замену, где в контексте выражения

счастья слово *gay* уступает место слову *happy* (Tahmasebi [et al.], 2021), (Periti [et al.], 2024).

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 (Tahmasebi [et al.], 2021), (Майсак, 2016).

В рамках настоящей работы рассматриваются семантические изменения лексического значения слов, без затрагивания иных явлений.

1.2. Семантическое описание слов

И.А. Стернин выделяет следующие принципы, применение которых необходимо в практике семантического описания (Стернин, Рудакова, 2017).

- Принцип неединственности метаязыкового описания ментальных единиц: семантика ментальных единиц может описываться разными метаязыками, и различия в этих описаниях требуют анализа и унификации, а не считаются ошибками.
- Принцип дополнительности семантических описаний: разные семантические описания языковых единиц дополняют друг друга и могут быть объединены в обобщающее описание.
- Принцип дополнительности словарных дефиниций: разные дефиниции лексической единицы в словарях отражают различные аспекты значения, и наиболее полное описание достигается их интеграцией.
- Принцип денотативной дифференциации значений: каждому уникальному денотату, обозначаемому словом, соответствует свое значение.

Итак, для наиболее полного описания значения слова на основе данных из словарей необходимо обобщить информацию из нескольких слова-

рей. Для этого нужно собрать и объединить определения из различных словарей, относящихся к одному и тому же современному периоду, провести денотативную дифференциацию значений и описать смысловую структуру значений.

Так, алгоритм применения метода обобщения словарных дефиниций, по мнению И.А. Стернина, заключается в следующем (Стернин, Рудакова, 2017):

1. Выписываются значения слова из всех доступных словарей.
2. Составляется единый список значений слова из разных словарей.
3. Уточняется список значений по денотативному принципу. Если слово номинирует некий денотат, отличный от других денотатов, фиксируется отдельное значение.
4. Анализируются примеры из словарных статей. Формулируются новые значения, если они выявляются только из примеров.
5. Каждое значение представляется с дефинициями из разных словарей.
6. Формулируется новый более развёрнутый вариант дефиниции, если требуется точность. Например, вместо синонимического ряда «юрисконсульт, адвокат» необходимо обобщение значений синонимов и формулировка семемы: ’специалист, защищающий чьи-либо интересы в суде, оказывающий юридические консультации; то же, что юрисконсульт, адвокат (разг.)’
7. Обновляется состав семантемы при отсутствии некоторых значений в словарях.
8. Местоимения в метаязыковых обозначениях заменяются на архисемы для унификации.
9. Функциональные и стилистические пометы обобщаются в альтернативной форме.

10. Актуализируются функциональные пометы, если они не соответствуют современному употреблению.
11. Если значение устарело, добавляется помета <<устар>>.
12. Территориальные семы обобщаются пометой <<обл.>> при указании конкретного региона.
13. Приводится совокупность примеров употребления слова из разных словарей.
14. Упорядочиваются значения многозначного слова от ядерных к периферийным.
15. Все значения приводятся в обобщенном виде с одним примером употребления каждый.

1.3. Обзор существующих методов

1.3.1. Историческая справка

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 (Виноградов, Шведова, 1999), а также книгу «Два века в двадцати словах», в деталях описывающую историю значения двадцати слов (Данова [et al.], 2018).

Книга «Два века в двадцати словах» представляет собой исследование, посвященное изменениям значений 20 интересных с точки зрения их эволюции слов в русском языке на протяжении XIX и XX веков (Данова [et al.], 2018). Ее создание стало возможным благодаря использованию Национального корпуса русского языка (НКРЯ), который является огромным электронным хранилищем текстов с начала XVI века до наших дней.

Хотя ручные методы продолжают применяться в лингвистике, появление цифровых корпусов и диахронических текстовых баз данных открыло новые перспективы для исследований. Цифровизация документов в различ-

ных областях не только облегчила доступ к текстам, но и позволила разработать полуавтоматические и автоматические методы анализа. Эти методы способны значительно расширить и углубить исследования изменений семантики слов, а также упростить их проведение (Tahmasebi [et al.], 2021).

1.3.2. Определение эмбеддингов

Множество исследований по автоматическому анализу семантических изменений обращалось к векторным представлениям слов – эмбеддингам, или вложениям. Они представляют собой метод преобразования слов в численные векторы фиксированной размерности. Это позволяет моделям машинного обучения работать с текстом, который изначально представлен в виде строк, переведённых в числовую форму. Главное их преимущество в том, что близкие по смыслу слова получают близкие векторные представления (Jatnika [et al.], 2019).

Одним из наиболее популярных методов для создания эмбеддингов является алгоритм Word2Vec, предложенный командой Google (Mikolov [et al.], 2013).

Word2Vec предлагает два основных подхода – Continuous Bag of Words (CBOW) и Skip-Gram. CBOW предсказывает текущее слово по окружающим, тогда как Skip-Gram предсказывает окружающие слова по текущему слову.

1.3.3. Метрики для измерения разницы между эмбеддингами

Для сравнения эмбеддингов используются различные метрики, которые измеряют степень их сходства или различия:

- Косинусное расстояние: Одна из наиболее популярных метрик, измеряющая косинус угла между двумя эмбеддингами. Если эмбеддинги направлены в одну сторону, косинусное расстояние близко к 0, если в противоположные – к 2.

$$d_{\cos}(A1, B1) = 1 - \cos(\theta) = 1 - \frac{A \cdot B}{\|A\|\|B\|} \tag{1}$$

где $A1$ и $B1$ – это эмбеддинги.

- Евклидово расстояние: Измеряет «прямую» дистанцию между двумя точками в пространстве.

$$d_{\text{e}}(A1, B1) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2} \tag{2}$$

где $A1$ и $B1$ – это эмбеддинги.

1.3.4. Кластеризация эмбеддингов

Эмбеддинги также можно кластеризовать, группируя их по схожести. Это позволяет, например, выявлять схожие группы слов или документов.

Наиболее популярные алгоритмы кластеризации включают:

- K-средних

  Алгоритм, который делит данные на K кластеров, минимизируя внутрикластерное расстояние. Данный алгоритм способен находить заранее установленное пользователем число кластеров.

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise) (Ester [et al.], 1996)

  Группирует точки по плотности, что позволяет работать с кластерами произвольной формы и игнорировать шум. Кроме того, DBSCAN способен выявлять неопределённое количество кластеров.

Основными типами векторных представлений, используемых для изучения семантических изменений, являются статические и контекстуальные эмбеддинги, которые мы рассмотрим далее.

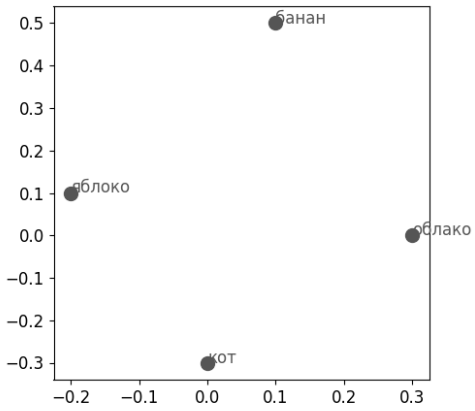
1.3.5. Статические эмбеддинги

Статические эмбеддинги дают представление слова для всего корпуса, на котором модель была обучена (Tahmasebi [et al.], 2021).

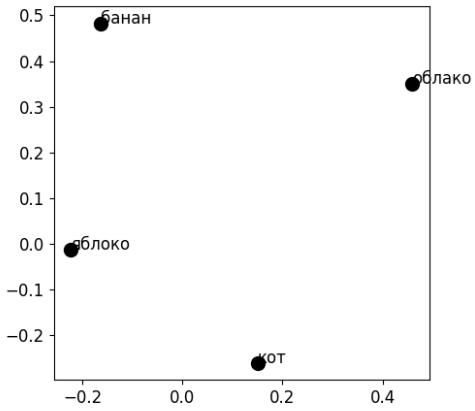
В качестве примера работы, использующей статические эмбеддинги на материале русского языка, можно привести проект Shiftry (Kutuzov, Fomin

[et al.], 2020), в котором для анализа семантических сдвигов использовались модели Word2Vec (Jatnika [et al.], 2019). Эти модели были обучены на обширном корпусе русскоязычных новостных текстов, охватывающем период с 2010 по 2020 годы, и позволили отследить диахронические изменения в употреблении слов. Поскольку при использовании статических эмбеддингов возможно производить только один вектор для одного слова в одном корпусе, корпус текстов был разделён по годам, позволив производить отдельные вектора слов для каждого года. В этом случае проявляется проблема. Хотя относительное положение эмбеддингов относительно друг друга для разных лет может быть сохранено, модели обучаются отдельно друг от друга, и их векторные пространства находится в разных системах координат. Поэтому, чтобы сравнивать эмбеддинги из разных периодов на значимом уровне, необходимо выровнять векторные пространства. Для этого используется метод Прокруста, который помогает привести векторы к общей системе координат, далее для подсчета степени семантического сдвига используется косинусное расстояние между векторами из различных временных срезов. За некоторыми исключениями, исследования с использованием статических векторов придерживаются аналогичной проекту Shiftry методологии (Tahmasebi [et al.], 2021).

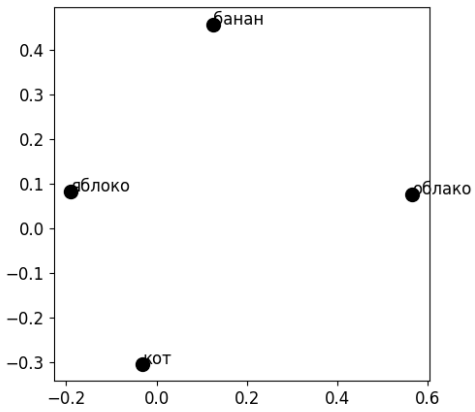
Пример работы вявления семантического изменения слова *облако* с использованием статических эмбеддингов и метода Прокруста вы можете увидеть на Рисунке 1.



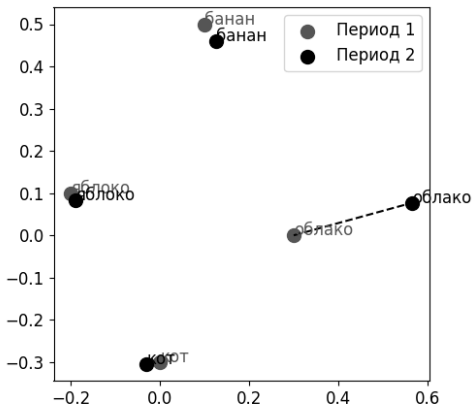
(а) Эмбеддинги для слов из эпохи 1.



(б) Эмбеддинги для слов из эпохи 2.



(в) Эмбеддинги эпохи 2 после применения метода Прокруста.



(г) Сравнение векторных пространств. Выявление сдвига слова *облако*.

Рис. 1. Пример поиска семантических изменений для статических эмбеддингов с использованием метода Прокруста.

Статические эмбеддинги оставались наиболее актуальными в задаче определения семантических изменений вплоть до 2020, где показывали лучшие результаты в соревновании по обнаружению изменений значения слов SemEval-2020 Task 1 (Schlechtweg [et al.], 2020).

Среди недостатков статических эмбеддингов 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,
- моделируется только усреднённое значение слова на основе его употребления в корпусе, не позволяя различать отдельные значения слова.

1.3.6. Определение контекстуальных эмбеддингов

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 присваивают токенам (минимальным единицам текста, с которыми работают модели, обычно это слова, части слова или пунктуация) различные эмбеддинги в зависимости от их контекста, что позволяет различать отдельные значения одного слова.

Например, в работе А.В. Кутузова приводится наглядный пример работы контекстуальных эмбеддингов (Kutuzov, 2020). На Рисунке 2 показана проекция вложений ELMo для слова *cell* (клетка) в 2000 годы в английском языке. На визуализации видны три кластера, отражающие различные значения слова *cell* (клетка). Два кластера, расположенные слева, представляют традиционные значения: внизу биологическое значение, связанное с клетками живых организмов, и вверху тюремное, где *cell* означает камеру. Кластер, который занимает правую сторону рисунка и чётко отделён от левых, демонстрирует современное употребление слова *cell*, где оно используется в контексте мобильной связи.

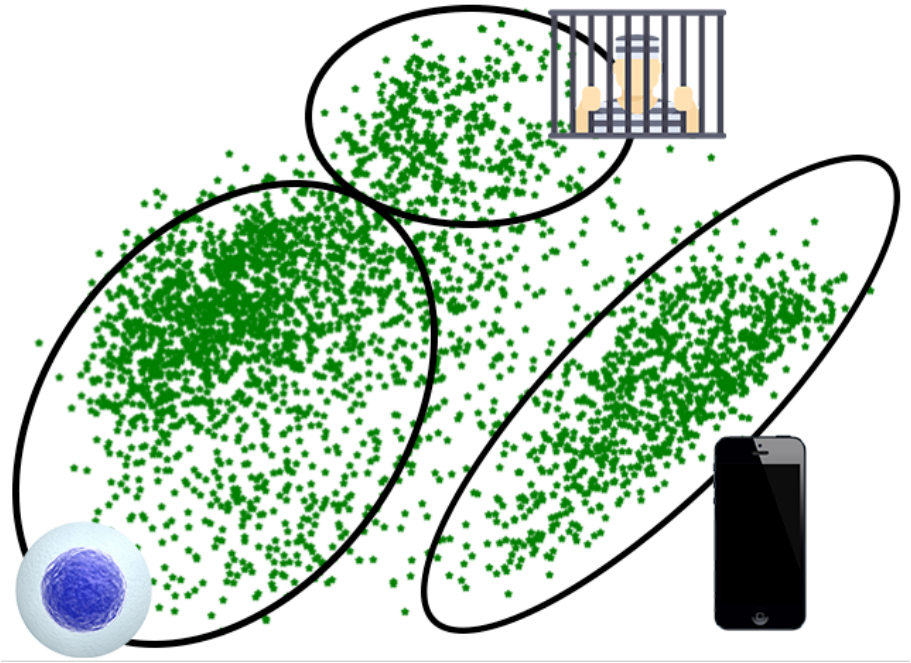


Рис. 2. Проекция эмбеддингов использований слова *cell* в 2000-ые годы

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 (Rodina [et al.], 2020).

1.3.7. Соревнование по выявлению семантических изменений Rushifteval

Теме автоматического выявления семантических сдвигов для русского языка было посвящено соревнование RuShiftEval, прошедшее в 2021 году (Kutuzov, Pivovarova, 2021). 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 предлагалась задача, которую можно свести к оценке семантической связи между значениями целевого слова в парах предложениях из разных временных периодов. Оценки (от 1 до 4) отражают степень семантического родства между значениями слова, где 1 обозначает отсутствие связи между значениями, а 4 – их совпадение. Затем индивидуальные оценки усредняются, формируя общую меру семантической родственности между употреблениями слова в разные временные периоды.

Вхождение в датасет Rushifteval представлено в Таблице **??**.

Таблица 1. Пример вхождения в датасет Rushifteval

Слово	радикал
Предложение 1	А вот социалисты и наши ***радикалы*** -- это совсем другого подбора.
Предложение 2	При некоторых условиях при отдельных элементарных реакциях возникают сразу два ***радикала,*** что приводит к разветвлению цепи.
Среднее	1.0
Аннотатор 1	1
Аннотатор 2	1
Аннотатор 3	1

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: RuShiftEval-1 (изменение значений между досоветским и советским периодом), RuShiftEval-2 (изменение значений

между советским и постсоветским периодом) и RuShiftEval-3 (изменение значений между досоветским и постсоветским периодом). В качестве метрики оценки использовалась ранговая корреляция Спирмена между ранжированием слов, сгенерированным системой, и эталонным ранжированием, полученным в ручной аннотации. После этого бралась средняя оценка между ранжированиями.

Пример того, как выглядит финальный файл с оценками, представлен в Таблице 2.

Таблица 2. Отрывок тестового файла, содержащего обобщённые оценки аннотаторов

слово	**досоветский: советский**	**советский: постсоветский**	**досоветский: постсоветский**
авторитет	3.233	2.956	2.844
амбиция	3.111	3.444	3.333
апостол	3.494	3.427	3.424
благодарность	3.233	3.567	3.656

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 (Rachinskiy, Arefyev, 2021). Чтобы решить это, они дообучали модель XLM-R на задаче генерации эмбеддингов, максимально близким к таким, какие получены на соответстующим использованиям слов словарным определениям (Conneau [et al.], 2019).

При дообучении их система включает в себя два отдельных кодировщика на основе XLM-R: Кодировщик контекстов для кодирования предложения с целевым словом и кодировщик глоссов для кодирования определения слова. Система оценивает возможные значения смысла слова путём сравне-

ния 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, то есть к значениям от 1 до 4.

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*.

Второй подход к решению задачи определения семантических изменений был также получен в рамках соревнования RuShiftEval и представлен в (Arefyev [et al.], 2021).

Исследователи обучали модель XLM-R на обширном многоязычном датасете Word-in-Context, а затем дообучали ее на наборе данных RuSemShift (аналог Rushifteval с другими лексемами) для настоящей задачи.

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

Так, применимость таких методов была подвергнута сомнению в работе (Giulianelli [et al.], 2023), где утверждается, 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 (Rachinskiy, Arefyev, 2021), где исследователи предлагают способ визуализации и интерпретации результатов. Однако у этого метода есть свои недостатки, обсуждаемые выше, а также необходимость опреде-

лять заранее значения, для которых будет строиться визуализация. Учитывая эти факты, новый подход, включающий моделирование определений, вызывает интерес для задачи обнаружения семантических изменений.

1.3.8. Моделирование определений

Моделирование определений (также генерация определений) описывается исследователями как «задача генерации определений слов в формате, который может быть прочитан людьми, как те, что можно найти в словарях», где на вход подается целевое слово и пример его использования, а на выход ожидается сгенерированное определение (Giulianelli [et al.], 2023). Вы можете увидеть пример в Таблице 3.

Таблица 3. Пример моделирования определений

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

Начало интереса к моделированию определений как теме исследования в области обработки естественного языка можно отнести к работе Noraset et al. (Noraset [et al.], 2016). В работе они исследовали 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.], 2018) выделили важное условие для моделирования

определений: необходимость контекста для точного захвата нюансов языка. Было предложено включить 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. Однако, у данных подходов есть следующие недостатки (Huang [et al.], 2021):

- проблема слов вне словаря, когда модели сталкиваются с трудностями в работе со словами, не встречавшимися во время обучения;
- проблемы избыточной и недостаточной специфичности в определениях.

Исследователи сообщают: «Избыточно специфичные определения представляют узкие значения слов, в то время как недостаточно специфичные определения представляют общие и нечувствительные к контексту значения.» В (Huang [et al.], 2021) предлагается использовать 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.

На момент написания работы, по мнению автора, лучший реализацией метода моделирования определений является решение, представленное в (Giulianelli [et al.], 2023).

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 ис-

пользуют модель Flan-T5, версию трансформера T5, большую генеративную языковую модель, дополнительно обученную на 1800 задачах по обработке естественного языка.

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 предлагают использовать 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 по схожей схеме с датасетом соревнования Rushifteval.

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 и эмбеддингами, полученными на основе них. Результаты показали, что сгенерированные определения обладают более высокой степенью корреляции с данными из DWUG, чем контекстуальные эмбеддинги.

Кроме того, исследователи обнаружили, что эмбеддинги определений образуют более плотные и четко определенные кластеры по сравнению с традиционными эмбеддингами, что делает их подходящими для представления значений слов.

Далее авторы исследовали возможность присваивать 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 такое, эмбеддинг которого наиболее близок к среднему значению всех эмбеддингов в кластере согласно скалярному произведению.

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.

1.4. Метрики оценки качества сгенерированных определений

При запуске модели на тестовой части датасета, представляется возможным сравнить каждое полученное определение с эталонным определением из датасета.

Одним из основных способов оценки сгенерированных определений являются метрики сходства строк (Gardner [et al.], 2022).

Примером такой метрики является BLEU (Bilingual Evaluation Understudy). BLEU — это стандартный алгоритм, используемый для оценки машинных переводов (Papineni [et al.], 2002). BLEU рассчитывается как точность n-грамм, то есть отношение правильных n-грамм к общему числу n-грамм в выходной строке. Недостатком BLEU является то, что он оценивает только совпадение n-грамм.

$$\text{BLEU} = \exp\left(\text{BP} + \sum_{n=1}^{N} w_n \log p_n\right), \tag{3}$$

где BP — это brevity penalty, который вычисляется как

$$\text{BP} = \min\left(1 - \frac{L_r}{L_c}, 0\right), \tag{4}$$

где L_r — длина эталонного определения, L_c — длина сгенерированного определения, w_n — веса для n-грамм, $p_n = \frac{\text{Число пересечений n-грамм}}{\text{Общее число n-грамм в сгенерированном тексте}}$ — точность n-грамм,.

Например, для текстов *Кошка залезла на шкаф* и *Кошка залезла на стол* будет 59.46 (от 100).

Еще одной популярной метрикой является ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation). ROUGE измеряет совпадение n-грамм между эталонным и кандидатом на определение (Lin, 2004). ROUGE-L — это модифицированная версия ROUGE, которая использует наибольшую общую подпоследовательность для измерения сходства между двумя определениями. Преимуществом ROUGE-L является то, что он автоматически определяет самые длинные последовательные общие n-граммы. Формула для расчета ROUGE-L выглядит следующим образом:

$$\text{ROUGE-L} = \frac{LCS(X, Y)}{L_r} \tag{5}$$

где $LCS(X, Y)$ — длина наибольшей общей подпоследовательности между строками X и Y, а L_r — длина эталонного определения.

Например, для *Кошка залезла на шкаф* и *Кошка залезла на стол* ROUGE-L составит 0.75 (от 1).

Тем не менее, у таких метрик есть недостаток. Они анализируют не семантику слов, а только буквальное совпадение n-грамм между ними.

Примером более продвинутой метрики является BERTScore. BERTScore (Bidirectional Encoder Representations from Transformers) —

это метрика, которая вычисляет оценку сходства между кандидатом и эталонным определением на основе предварительно обученных контекстуальных эмбеддингов из BERT (Zhang [et al.], 2020). BERTScore вычисляет точность (6), полноту (7) и F1-меру (8). Формулы для расчета BERTScore выглядят следующим образом:

$$\text{Precision} = \frac{1}{|C|} \sum_{x \in C} \max_{y \in R} \cos(x, y) \tag{6}$$

$$\text{Recall} = \frac{1}{|R|} \sum_{y \in R} \max_{x \in C} \cos(x, y) \tag{7}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

где C — множество токенов кандидата, R — множество токенов эталона, и $\cos(x, y)$ — косинусное сходство между эмбеддингами токенов x и y.

Например, для текстов *выложил документ в удалённое хранилище*, *выгрузил файл в облако* значение BERT-F1 – 76.21 (от 100), несмотря на непосредственное совпадение лишь одного слова.

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 (Papineni [et al.], 2002), (Lin, 2004), (Zhang [et al.], 2020). Так, в обзорной статье по моделированию определений утверждается, что на момент выпуска статьи метрика BLUE использовалась в 9 научных публикациях по теме, а ROUGE-L и BERTScore – в 3 (Gardner [et al.], 2022).

1.5. Классификация ошибок сгенерированных определений

В работах (Huang [et al.], 2021) и (Noraset [et al.], 2016) проводится оценка выборки сгенерированных определений, для которой они предсталяют классификацию ошибок, часто встречаемых в сгенерированных определениях. Далее представлена классификация ошибок на основе классификации (Huang [et al.], 2021).

Исключением является «Избыточность или чрезмерное использование общих фраз», которую авторы опустили в исследовании из-за её редкости.

Таким образом, ошибками являются:

1. **Избыточное конкретизирование.** Определение даёт слишком узкое описание слова.

   Слово: дуновение

   Пример: дуновение – ’движение неприятного запаха через воздух’.

   Эталонное определение: «Лёгкий порыв ветра; движение воздуха.» (Кузнецов, 1998)

   Объяснение ошибки: Ошибка заключается в том, что определение ограничивает значение слова компонентом ’неприятный запах’, тогда как оно имеет более широкое значение.

2. **Недостаточное конкретизирование.** Определение даёт слишком общее или неполное описание.

   Слово: капитан

   Пример: капитан – ’член команды’.

   Эталонное определение: «Командир, начальник судна.» (Кузнецов, 1998).

   Объяснение ошибки: Ошибка заключается в отсутствии указания на ключевые семантические компоненты руководящей роли капитана, что приводит к недостаточной конкретизации значения слова.

3. **Самореференция.** Определение содержит само слово или его производные.

   Слово: самосознание

   Пример: самосознание – ’состояние, при котором у человека присутствует самосознание’.

   Эталонное определение: «Полное понимание самого себя, своего значения, роли в жизни, обществе.» (Кузнецов, 1998).

   Объяснение ошибки: Ошибка заключается в том, что в определении не раскрывается семантическая сущность слова, так как для описания значения используется само же слово.

4. **Неправильная часть речи.** Модель даёт такое определение, которое относится к лексеме другой части речи.

   Слово: стекло

   Пример: стекло – ’переместиться вниз, сбежать (о жидкости)’ для контекста «После урока стекло оказалось сломанным.»

   Эталонное определение: «изделие из твёрдого прозрачного материала.» (Ахапкин [et al.], 2003)

   Объяснение ошибки: Определение относится к глаголу *стекло*, тогда как слово в контексте является существительным.

5. **Противоположное значение.** Определение выражает смысл, противоположный истинному значению слова.

   Слово: внутрь

   Пример: внутрь – ’ненаправленный в центр’.

   Эталонное определение: «Во внутреннюю часть, в пределы, в глубину, в середину.» (Кузнецов, 1998).

   Объяснение ошибки: Определение выражает противоположное значение истинному значению слова.

6. **Близкая семантика.** Определение верно передает только часть сем.

Слово: машина

Пример: машина – ’устройство с автоматическими функциями’.

Эталонное определение: «Механизм или совокупность механизмов, совершающие какую-л. полезную работу путём преобразования одного вида энергии в другой.» (Кузнецов, 1998).

Объяснение ошибки: Определение верно передает только отдельные стороны денотата, но упускает ключевые.

7. **Некорректность.** Определение полностью ошибочно и не соответствует значению слова.

   Слово: первый

   Пример: первый – ’следующий после всех остальных в списке предметов’.

   Эталонное определение: «При счёте вы называете первым предмет, элемент, человека и т. д., с которого начинаете счёт.» (Ахапкин [et al.], 2003)

   Объяснение ошибки: Определение полностью ошибочно.

8. **Избыточность или чрезмерное использование общих фраз.** Определение содержит повторы или избыточные формулировки.

   Слово: спутник

   Пример 1: спутник – ’тот, кто совершает путь, путь вместе с кем-л.’.

   Эталонное определение: «Тот, кто совершает путь вместе с кем-л.» (Кузнецов, 1998).

   Объяснение ошибки: Определение содержит повтор слова *путь*.

Также среди определений можно выделить:

Корректные. Определение точно и в нужной степени полно передает значение слова и не содержит какие-либо из вышеописанных ошибок. Слово: винодельня Пример: винодельня – ’заведение, помещение для изготовления вина’. Эталонное определение: «Заведение,

место, где изготовляется виноградное вино.» (Ушаков, 1940) Объяснение: Определение точно и полно передает значение слова.

Выводы

В последние годы возрос интерес к полуавтоматическим и автоматическим подходам выявления семантических изменений, основанным на векторных представлениях слов (эмбеддингах). Статические эмбеддинги, обучаемые на корпусах текстов, позволяют выявлять семантические сдвиги, однако имеют ограничения, связанные с необходимостью выравнивания моделей и неразличением значений отдельного слова. Контекстуальные эмбеддинги, генерируемые современными языковыми моделями, показывают более высокую точность в задачах обнаружения семантических изменений, поскольку учитывают контекст употребления слов.

Новым перспективным направлением является использование моделирования определений слов на основе генеративных больших языковых моделей. Сгенерированные определения демонстрируют более высокую корреляцию с данными о семантической близости слов, чем традиционные эмбеддинги, и могут служить семантическим представлением слов, превосходящим по качеству векторные репрезентации. Существует несколько способов оценки сгенерированных определений: от метрик сходства текста до классификаций для качественного анализа.

Тем не менее, моделирование определений остаётся недостаточно исследованной темой в контексте семантических изменений, особенно на материале русского языка.

Глава 2. Предлагаемый подход

В данном исследовании предлагается метод моделирования определений слов, который характеризуется как задача генерации определений в формате, доступном для чтения людьми, аналогично тем, что представлены в словарях. Входными данными для модели служат целевое слово w и пример его использования в предложении s, на выходе генерируется определение d.

Основными этапами предлагаемого подхода являются:

Этап 1: Обучение

Пусть имеется датасет $D = \{(w_i, s_i, d_i)\}_{i=1}^{N}$, где w_i — это слово, s_i — пример его использования в предложении, а d_i — определение данного слова. Целью первого этапа является обучение генеративной большой языковой модели. Модель M обучается на данных словаря, содержащих определения слов и их контекст использования, с целью генерации определений слов, аналогичных тем, что можно найти в словарях. Формально, модель M обучается минимизировать функцию потерь L, определенную как:

$$L(M) = \sum_{i=1}^{N} \text{loss}(M(w_i, s_i), d_i), \tag{9}$$

где loss — это функция потерь, измеряющая расхождение между сгенерированным определением и эталонным определением.

Этап 2: Тестирование

На втором этапе проводится тестирование обученной модели по ряду метрик. Этот этап включает:

1. Генерацию определений для тестовой выборки $D_{\text{test}} = \{(w_j, s_j, d_j)\}_{j=1}^{M}$ и последующую оценку качества сгенерированных определений $\hat{d}_j = M(w_j, s_j)$ относительно эталонных определений d_j. Для оценки качества используются меры сходства текста, которые оценивают

формальную схожесть или семантику сравниваемых текстов. Формально, метрика metric определяется как:

$$\text{metric} = \frac{1}{M} \sum_{j=1}^{M} \text{similarity}(\hat{d}_j, d_j) \tag{10}$$

где similarity — такая функция, измеряющая сходство между сгенерированным и эталонным определениями, как BLEU, ROGUE-L и BERT-F1.

2. Проверку модели на датасете, посвященном детектированию семантических изменений и содержащих оценки аннотаторов между парами вхождений.

Для каждой пары использований слов в датасете генерируются определения $(\hat{d}_{k1}, \hat{d}_{k2})$, которые затем векторизуются $(\vec{d}_{k1}, \vec{d}_{k2})$. Полученное значение расстояния между векторизованными определениями $\text{dist}(\vec{d}_{k1}, \vec{d}_{k2})$ сравнивается с оценками аннотаторов в датасете.

Этап 3: Визуализация

Для визуализации семантических изменений слов полученные с помощью модели определения векторизуются с помощью векторизатора V.

$$\mathbf{v}_i = V(d_i) \tag{11}$$

Определения, имеющие семантически близкие значения, группируются с использованием алгоритма кластеризации C.

$$\{K_1, K_2, \ldots, K_m\} = C(\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}), \tag{12}$$

где m – количество кластеров.

Для каждого кластера K_j выбирается прототипическое определение $\hat{d}_{\text{proto}}$, которое определяется как наиболее близкое к центру кластера (центроиду).

Пусть $\mathbf{c}_j$ — центроид кластера K_j:

$$\hat{d}_{\text{proto},j} = \arg\min_{\mathbf{v} \in K_j} d(\mathbf{v}, \mathbf{c}_j)$$

,

где d — метрика расстояния.

Затем создаются столбчатые диаграммы, отражающие частоту употреблений различных значений слова во времени, с обозначением категорий цветовой градацией и легендой.

Этап 4: Качественный анализ

Проводится качественный анализ результатов, полученных с помощью визуализаций, на наборе слов. Полученные определения сравниваются с таковыми из словарей, а статистическая информация о частоте использования тех или иных значений сравнивается с информацией из лексикографических изданий.

Глава 3. Анализ предлагаемого подхода

3.1. 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 (Zmitrovich [et al.], 2023). Для выбора модели использовался бенчмарк для оценки продвинутого понимания русского языка RussianSuperGLUE (Shavrina [et al.], 2020). В бенчмарке присутствуют несколько 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. На момент написания работы 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;
- продолжение текстовых последовательностей.

Каждый из денойзеров помечен специальным токеном. В настоящей работе при работе с моделью используется денойзер, помеченный специальным токеном «<LM>», который задействован в задаче продолжения текста.

В качестве материала, используемого для обучения модели, выступил датасет, включающий в себя материал из МАС – «Малого академического словаря» (Евгеньева, 1981-1984). Материал МАС 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. Пример вхождения в набор данных представлен в Таблице 4. Определения выделены жирным.

Таблица 4. Информация о лексеме из МАС

Лексема	**Определения и примеры использования**
производительность	**Способность производить, выпускать то или иное количество продукции:** Производительность машин. Годовая производительность завода. **Целесообразность, плодотворность:** Производительность затрат.

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*. В результате, было получено 122 тысяч 350 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. Та-

ким образом, на вход модель принимает лексему и контекст, в которой она употреблялась, а на выход ожидается сгенерированное определение.

Таблица 5. Пример отформатированного запроса модели

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

FRED-T5-1.7B была дообучена на полученном из «Малого академического словаря» материале в течение 6 эпох с линейным шагом обучения 0.001, размером батча 16 и оптимизатором Adafactor на машине с одной видеокартой RTX 3090 с объёмом видеопамяти 24ГБ. Обучение заняло 9 часов 40 минут. Максимальное использование видеопамяти не превышало 13ГБ. 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), сэкономить используемую память и ускорить обучение. В качестве метрики потерь (лосса) используется кросс-энтропия. График потерь представлен на Рисунке 3.

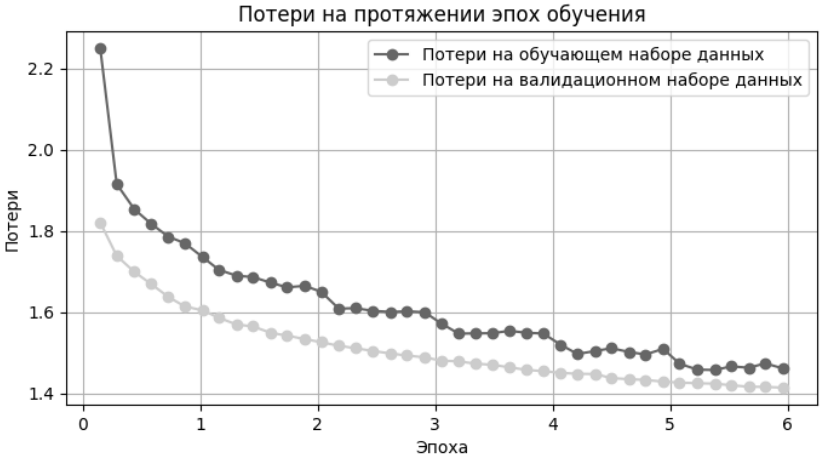


Рис. 3. Потери при обучении модели

Более подробный обзор гиперпараметров модели доступен в Приложении Б.

3.2. Тестирование модели метриками сходства строк

В данной работе использовались версии метрик BLEU, ROUGE-L и BERT-F1, взятые из библиотеки *evaluate* (URL: https : / / github . com / huggingface/evaluate). Результаты оценки дообученной модели представлены в Таблице 6.

Таблица 6. Результаты дообучения FRED-T5-1.7B на датасете MAC (от 100, больше – лучше)

Метрика	Значение
BLEU	11.02
ROUGE-L	29.36
BERT-F1	75.22

Тестирование показывает низкие значения метрик BLEU и ROUGE-L, что говорит о том, что модель формулирует определения не так, как они написаны в тестовой выборке. Тем не менее, это не означает некорректность генерируемых определений. Судя по результатам метрики BERT-F1, можно сказать, что семантически сгенерированные определения совпадают с таковыми из тестовой выборки. Такие результаты можно объяснить тем, что модель выдаёт семантически верные, но иначе сформулированные определения. Например, для слова *отощалый* в контексте «Иван Бедный сидел в развалившейся лачуге, худой, отощалый.» ожидаемым определением являлось ’ставший тощим, отощавший’, но моделью было сгенерировано ’сильно исхудавший от недоедания’. Оба данных определения корректны и описывают того, кто стал худым, но при этом между ними не совпадает ни единого токена. Метрики BLEU и ROUGE-L для такой пары показывают 0. Только BERT-F1 возвращает 70.39, поскольку данная метрика вместо совпадения после-

довательности сравниваемых определений использует векторизатор и сравнивает расстояние между полученными эмбеддингами, которое обозначает семантическую близость двух текстов.

Следует сказать, что предварительный анализ полученных определений показал наличие определений, имеющих ошибку самореференции. Например, для контекста *[Пантелей Прокофьевич] ничего не мог сделать, чтобы восстановить в семье прежний порядок.* и целевого слова *восстановить* изначально было сгенерировано «привести в прежнее состояние; восстановить», что содержит повторение целевого слова, а результаты метрик были ниже представленного ранее: 10.64 для BLEU, 29.09 для ROUGE-L, и 75.19 для BERT-F1.

Удалось значительно сократить количество определений с самореференцией, исключив из генерации токены, соответствующие целевому слову. Для этого мы заранее определяли все возможные формы целевого слова и исключали их из процесса создания определения, что помогло избавиться от подобных ошибок и улучшить результаты.

Например, для контекста *[Пантелей Прокофьевич] ничего не мог сделать, чтобы восстановить в семье прежний порядок.* и целевого слова *восстановить* после правки было сгененировано «привести в прежнее состояние, положение».

3.3. Тестирование модели на материале соревнования Rushifteval

С помощью дообученной модели FRED-T5-1.7B были получены определения для тестовой части датасета соревнования Rushifteval.

Для векторизации сгенерированных определений использовалась модель paraphrase-multilingual-mpnet-base-v2 (URL: https : / / huggingface . co / sentence-transformers/paraphrase-multilingual-mpnet-base-v2). Данная модель является одним из лидеров по задаче семантической схожести текстов (URL:

https://github.com/avidale/encodechka). Далее 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. Так, значения косинусного расстояния от 0 до 2, где 0 обозначает идентичность векторов, а 2 – их противоположность, преобразованы в значения от 1 до 4, соответствующие оценкам аннотаторов, где 1 – противоположные значения, а 4 – идентичные. Результаты представлены в Таблице 7.

Таблица 7. Коэффициэнты корреляции с использованием LinReg

Пары периодов	**Коэффициент корреляции**
Среднее	0.7156
досоветский:советский	0.7056
советский:постсоветский	0.7251
досоветский:постсоветский	0.7160

В качестве попытки улучшить результаты в течение трёх эпох был дообучен векторизатор paraphrase-multilingual-mpnet-base-v2 на материале RuSemShift (аналог RuShiftEval с другими лексемами). Результаты представлены в Таблице 8. Авторами соревнования рекомендуется дообучать решения на данном наборе данных для улучшения результатов. Параметры дообучения векторизатора доступны в Приложении В.

Таблица 8. Коэффициэнты корреляции с использованием LinReg и дообученного векторизатора

Пары периодов	**Коэффициент корреляции**
Среднее	0.8002
досоветский:советский	0.7843
советский:постсоветский	0.8139
досоветский:постсоветский	0.8023

Таким образом, благодаря дообучению векторизатор была улучшена производительность алгоритма на более чем 8%.

3.4. Сравнение с существующими подходами

Рассмотрим полученные результаты в сравнении с аналогами из соревнования Rushifteval.

Исходный код лидирующих решений выложен в открытый доступ и доступен для воспроизводства.

Авторы статьи (Rachinskiy, Arefyev, 2021) предоставляют доступ к части исходного кода их исследования (URL: https://github.com/myrachins/RuShiftEval). Так, были опубликованы следующие компоненты:

1. программный модуль генерации прогнозов на основе заранее вычисленных эмбеддингов, полученных с использованием обученной нейросетевой модели,
2. программный модуль оценки результатов.

Результаты исследования были воспроизведены и представлены в Таблице 9.

Таблица 9. Коэффициэнты корреляции

Пары периодов	Коэффициент корреляции
Среднее	0.8021
досоветский:советский	0.7808
советский:постсоветский	0.8032
досоветский:постсоветский	0.8223

Другой командой, представившей решение в открытый доступ является (Arefyev [et al.], 2021).

В Таблице 10 представлен воспроизведённый нами результат (URL: https://github.com/Daniil153/DeepMistake).

Таблица 10. Коэффициэнты корреляции

Пары периодов	Коэффициент корреляции
Среднее	0.8494
досоветский:советский	0.8563
советский:постсоветский	0.841
досоветский:постсоветский	0.8511

В Таблице 11 представлены результаты подхода в сравнении со всеми решениями, участвовашими в соревновании.

Таблица 11. Результаты алгоритма в сравнении с результатами команд Rushifteval.

Команда	**Среднее**	**Тип представления слов**	**Используемая модель**
DeepMistake (после соревнования)	**0.850**	контекст. эбм.	**XLM-R**
GlossReader	0.802	контекст.	XLM-R
Настоящий подход с дообучением векторизатора	**0.800**	**сген. опр.**	**FRED-T5-1.7B**
DeepMistake	0.791	контекст. эбм.	XLM-R
vanyatko	0.720	контекст. эбм.	RuBERT
Настоящий подход	**0.716**	**сген. опр.**	**FRED-T5-1.7B**
aryzhova	0.457	контекст. эбм.	RuBERT, ELMo
Discovery	0.453	контекст. эбм.	BERT
UWB	0.417	статич. эбм.	FastText
dschlechtweg	0.392	статич. эбм.	Word2Vec
jenskaiser	0.382	статич. эбм.	Word2Vec
SBX-HY	0.369	статич. эбм.	Word2Vec
Baseline	0.332	статич. эбм.	Word2Vec
svart	0.262	статич. эбм.	Word2Vec
BykovDmitrii	0.261	контекст. эбм.	XLM-R
fdzr	0.178	статич. эбм.	Word2Vec

Как видно из Таблицы 11, настоящее решение лучше по качеству большинства аналогов из соревнования Rushifteval. Два решения с самым высоким качеством были описаны ранее в Главе 1.

3.5. Визуализация результатов работы модели

Для создания визуализаций семантических изменений слов используются используются библиотеки *matplotlib* и *scikit-learn*. Полученные с помощью модели определения векторизуются с помощью дообученного на материале Rusemshift в прошлой главе векторизатора. Так как для слов, имеющих одинаковое значение, модель склонна генерировать семантически 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. Для достижения оптимальной кластеризации выбираются небольшие значения параметров и после повышаются, пока близкие определения, сформулированные похожим, но разным образом, не объединятся 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 рисунок, где представлена столбчатая диаграмма, показывающая процентное соотношение значений исследуемого слова за разные периоды времени. Каждая категория обозначена на диаграмме своим оттенком цвета и соответствующим временным интервалом. Под диаграммой находится расшифровка значений, а также использованные параметры визуализации.

Из достоинств такого подхода к визуализации на основе сгенерированных определений можно отметить способность выявлять произвольные значения вместо заранее определённых исследователем, например в отличие от подхода (Rachinskiy, Arefyev, 2021), что может быть полезно для исследования новых значений, которые ещё не задокументированы, а также тех значений, которые могут быть по каким-либо причинам не отражены в словарях.

Пример визуализации распроложен на Рисунке 4.

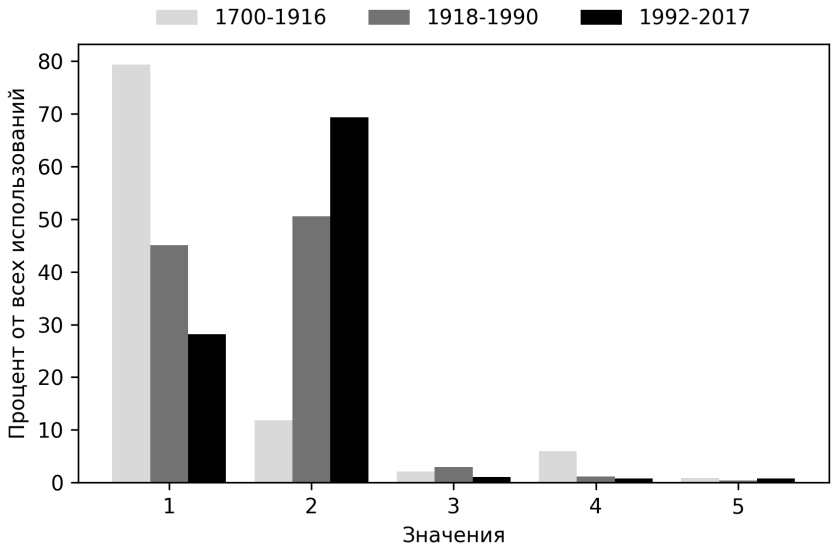


Рис. 4. Изменение значений слова *машина*

Значения для визуализации слова *машина* (Параметры: eps=0.14, min_samples=5).

1. Приспособление, устройство, служащее для выполнения какой-л. работы.

2. Автомобиль, транспортное средство.
3. Самолет, вертолет и т. п.
4. О человеке, действующем механически, бездумно.
5. Система, совокупность каких-либо учреждений, организаций, предприятий и т. п.

3.6. Код работы

Код, использованный во время выполнения настоящей работы, выложен в открытый доступ на сайте GitHub. (URL: https://github.com/tatarinovst2/work-definition-modeling)

Проект включает в себя 3277 строки кода на языке python3.11, 481 строку скриптов на языке bash, а также 562 строки документации.

Проект включает следующие модули:

- config: Запуск тестов и проверок проекта с помощью CI
- latex: Написание текста работы в формате LaTeX
- model: Обучение модели и тестирование метриками
- mas_parser: Краулер и парсер словаря МАС
- vizvector: Векторизация определений и визуализация результатов
- rushifteval: Тестирование алгоритма на материале соревнования Rushifteval
- ruscorpora: Позволяет осуществить качественный анализ алгоритма

Наличие тестов и открытого доступа к коду проекта позволяет сделать исследование воспроизводимым, а также готовым к внедрению в другие проекты.

3.7. Качественный анализ результатов работы алгоритма

Для дальнейшего анализа результатов алгоритма используются 20 слов с изменившемся значением из книги «Два века в двадцати словах» (Данова [et al.], 2018): *знатный*, *кануть*, *классный*, *мама*, *машина*, *молодец*, *пакет*,

передовой, *пионер*, *пожалуй*, *пока*, *привет пружина*, *публика*, *свалка*, *сволочь*, *стиль*, *тётка*, *тройка*, *червяк*. Использования данных слов взяты из диахронического корпуса НКРЯ.

Для каждого рассматриваемого слова из каждого периода (досоветсткий, советский и постсоветсткий) берётся выборка из 300 вхождений, где для каждого использования слова генерируется определение, а после строится соответствующая визуализация.

Далее для каждого слова описана семантика слов на основе словарей в соответствии с рекомендациями издания И.А. Стернина (Стернин, Рудакова, 2017). В соответствии с ними, так как невозможно построить полное описание значения слова с использованием только одного словаря, требуется обобщение данных нескольких словарей. В качестве материала взяты три словаря современного русского языка «Большой толковый словарь» (*БТС*) (Кузнецов, 1998), «Толковый словарь русского языка Дмитриева» (*ТСД*) (Ахапкин [et al.], 2003) и «Толковый словарь русского языка» Ожегова и Шведовой (*ТСО*), а также книга «Два века в двадцати словах». Книга «Два века в двадцати словах» использована при обобщении значений, поскольку наряду со словарями содержит описания значений. Исключением при обобщении значения является информация из помет, так как модель не обучалась на их генерацию. После чего проведено сравнение выявленных при семантическом описании лексемы значений и тех, что выявлены алгоритмом визуализации.

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.

Определения оцениваются по классификации, описанной в Главе 1 и построенной на основе работ (Huang [et al.], 2021) и (Noraset [et al.], 2016).

Далее представлен разбор результата работы предлагаемого подхода для 3 слов, чьи результаты анализа значительно отличаются друг от друга. После этого представлены общие выводы по качественному анализу.

Подробный анализ результатов по каждому из остальных 17 слов представлен в Приложении А.

Анализ слова *тройка*

В результате анализа семем лексемы *тройка* в толковых словарях были выделены девять значений, которые можно сформулировать следующим образом:

1. Количество три. *Тройка поплавков.* (*«Количество три.»* в БТС, *«Тройное количество чего-либо.»* в ТСД, *«(о сходных или однородных предметах) количество три.»* в ТСО, *«Количество, сумма из трех единиц.»* в «Двух веках в двадцати словах»)
2. Оценка успеваемости в пятибалльной системе, означающая удовлетворительно. *Учиться на тройки.* (*«Оценка успеваемости в пятибалльной системе, означающая удовлетворительно.»* в БТС, *«Школьная учебная отметка «удовлетворительно»».* в ТСО, *«В пятибалльной системе тройкой называют удовлетворительную, посредственную оценку чьих-либо знаний.»* в ТСД, *«Оценка в учебе.»* в «Двух веках в двадцати словах»)
3. Упряжка в три лошади. *Русская тройка.* (*«Три лошади в одной упряжке.»* в БТС, *«Упряжка в три лошади.»* в ТСО, *«Тройкой называют упряжку из трёх лошадей — коренной и двух пристяжных.»* в ТСД, *«Лошади.»* в «Двух веках в двадцати словах»)
4. Костюм, состоящий из пиджака (или жакета), брюк (или юбки) и жилета. *Прийти в тройке.* (*«Костюм, состоящий из пиджака (или*

жакета), брюк (или юбки) и жилета» в БТС, *«Костюм, состоящий из пиджака, брюк (или жакета, юбки) и жилета»* в ТСО, *«Костюм, который состоит из пиджака (или жакета), брюк (или юбки) и жилета»* в ТСД, *«Костюм»* в «Двух веках в двадцати словах»)

5. Игральная карта с тремя очками. *Играю тройками.* (*«Игральная карта в три очка.»* в БТС, *«В картах тройкой называют игральную карту в три очка»* в ТСД, *«Игральная карта с тремя очками»* в «Двух веках в двадцати словах»)
6. Цифра 3. *Написать тройку.* (*«Цифра 3»* в БТС, *«Цифра 3»* в ТСО, *«Тройка — это цифра 3»* в ТСД)
7. Транспортное средство, обозначенное цифрой 3. *Тройка изменила маршрут.* (*«Название автобуса, трамвая, троллейбуса третьего маршрута»* в БТС, *«Название чего-н. (обычно транспортного средства), обозначенного цифрой 3»* в ТСО, *«Маршрут автобуса, трамвая, троллейбуса, который пронумерован цифрой 3»* в ТСД)
8. Группа из трёх человек, обычно неразлучная. *Тройка знатоков.* (*«Тройкой называют устойчивую группу людей.»* в ТСД, *«Три человека»* в «Двух веках в двадцати словах»)
9. Звено боевых истребителей. *Идёт в пикирование вторая тройка.* (*«Тройкой называют звено боевых истребителей»* в ТСД)

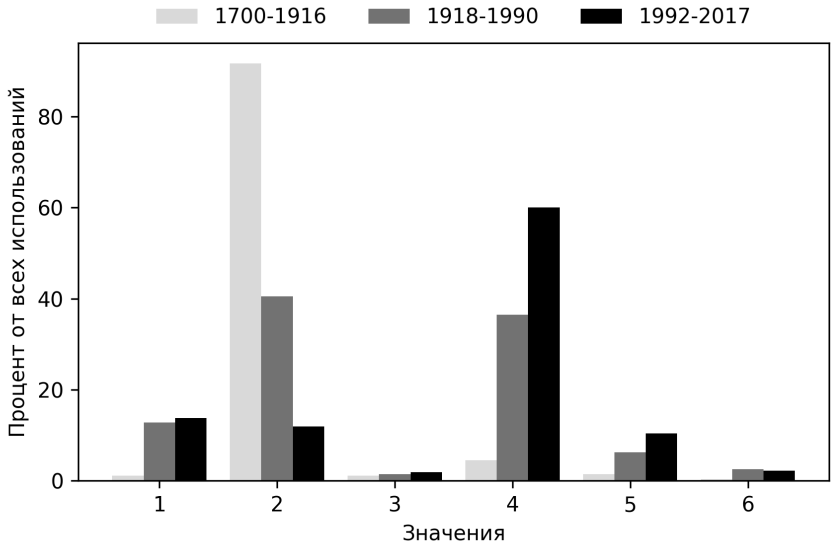


Рис. 5. Изменение значений слова *тройка*

Визуализация для слова *тройка* представлена на Рисунке 5. (Параметры: eps=0.28, min_samples=12) Соответствующие значения:

1. Неудовлетворительная оценка по какому-либо предмету.
2. Одна из трех лошадей, запряженных в такую повозку.
3. Игральная карта с тремя одинаковыми мастями.
4. Группа лиц, состоящая из трех лиц.
5. Число 3.
6. Старинная мужская верхняя одежда из сукна или бархата, застегивавшаяся спереди на пуговицы.

Анализ значений слова *тройка*

Пятое определение корректно сформулировано. Четвёртое определение избыточно. Первое, второе, третье и шестое определения не соответствуют обобщенным значениям.

- ’Число 3.’ полностью соответствует ’Число 3’.
- ’Неудовлетворительная оценка по какому-либо предмету.’ является близким значением, так как в обобщенных значениях имеется ’школьная оценка’ (’Оценка успеваемости в пятибалльной системе,

означающая удовлетворительно.’), но она означает «удовлетворительно», а не «неудовлетворительно».

- ’Одна из трех лошадей, запряженных в такую повозку.’ является близким обобщённому значению ’Упряжка в три лошади.’, но формулировка «одна из трех лошадей» некорректна, так как денотатом является упряжка, включающая в себя все три лошади и повозку.
- ’Игральная карта с тремя одинаковыми мастями.’ является близким значением, так как в обобщенных значениях имеется игральная карта с тремя очками, но формулировка «с тремя одинаковыми мастями» некорректна, мастью же является одна из четырёх категорий карт (В БТС: «Масть – один из четырёх разрядов на которые делится колода карт по цвету и форме очков.»).
- ’Группа лиц, состоящая из трех лиц.’ соответствует ’Группа из трёх человек, обычно неразлучная.’, так как включает те же семы *’группа лиц’*, *’три’*, однако содержит повторение слова *лиц*, которое делает его избыточным.
- ’Старинная мужская верхняя одежда из сукна или бархата, застегивавшаяся спереди на пуговицы.’ является некорректным определением, так как такое описание больше соответствует историческим видам одежды, таким как кафтан или сюртук, но не «костюму», представляющему собой комплект из брюк, пиджака и жилета.

Обобщённые значения, не найденные в визуализации:

- ’Цифра 3.’ отсутствует среди предложенных моделью значений. Можно предположить, что информации из контекста использований недостаточно для выделения этого значения.
- ’Транспортное средство, обозначенное цифрой 3.’ также отсутствует в визуализации. Однако, модель способна на выделение данного значения.

- ’Звено боевых истребителей.’ отсутствует среди предложенных моделью значений. Модель могла не выделить это значение из-за ограниченного количества использований.

Таким образом, для лексемы *тройка* представлено 1 корректное определение. Среди ошибок присутствуют:

- Избыточность или чрезмерное использование общих фраз: 1
- Некорректность: 1
- Близкое значение: 3

Перейдем к частотности значений.

Судя по книге «Два века в двадцати словах», данные которой представлены на Рисунках 6 и 7, в XIX веке для слова *тройка* преимущественным является значение ’Упряжка в три лошади.’ с около 95% использований, что также отражено в нашей визуализации с около 90% использований близкого значения ’Одна из трех лошадей, запряженных в такую повозку.’ Начало советского периода, утверждается в книге, характерно увеличением использования значения ’Три человека’ в 20-ые годы, а также ’Оценка успеваемости в пятибалльной системе, означающая удовлетворительно.’ в 40-ые годы. К концу советского периода используется множество значений и ’Упряжка в три лошади.’ больше не является лидирующим. Это информация отражена в нашей визуализации, где ’Одна из трех лошадей, запряженных в такую повозку.’ падает до 40% в советское время и 15% постсоветское. Вместо него самым частым становится ’Группа лиц, состоящая из трех лиц.’ с около 60% в постсоветский период, а также частыми становятся значения школьной оценки и числа 3 с 10-15% использования.

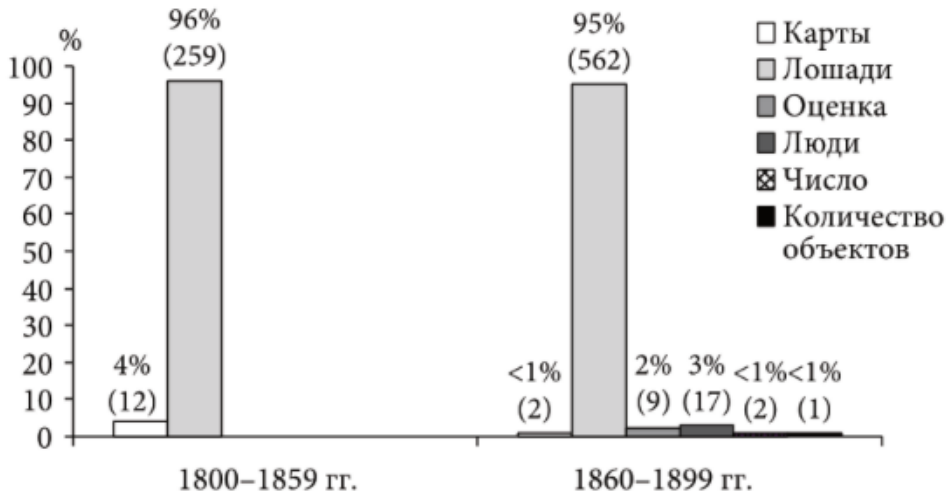


Рис. 6. Значения слова *тройка* для 1800-1899 согласно (Данова [et al.], 2018)

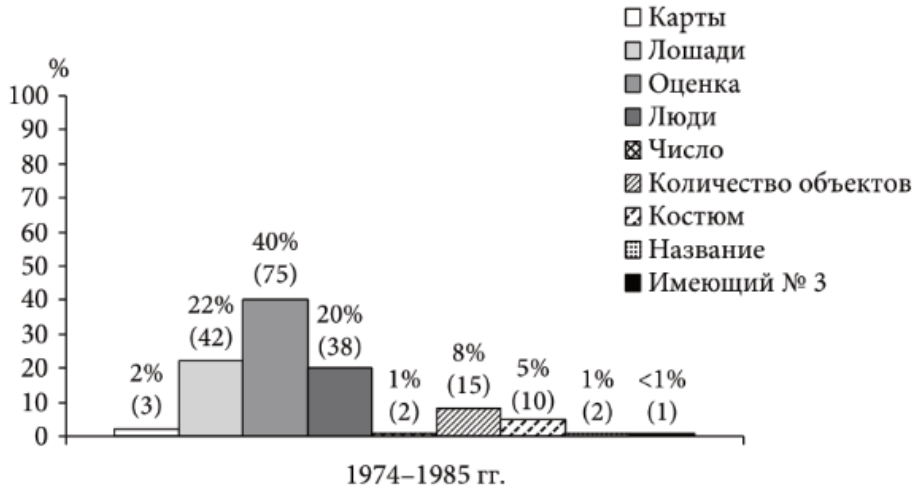


Рис. 7. Значения слова *тройка* для 1974-1985 согласно (Данова [et al.], 2018)

Таким образом, предложенный подход по большей части отражает значения, в которых использовалось слово *тройка*, не полностью согласуясь с данными из толкового словаря, но соответствуя историческим исследованием.

Анализ значений слова *пока*

1. В течение некоторого времени; до сих пор ещё; впредь до чего-л. *Пока ничего не известно.* (*«В течение некоторого времени; до сих пор ещё; впредь до чего-л.»* в БТС, *«В течение нек-рого времени, впредь до чего-н.; до сих пор еще.»* в ТСО, *«Наречие – в течение*

некоторого времени, до сих пор еще.» в «Двух веках в двадцати словах»)

2. В то время как. *Пока он учится, надо ему помочь.* (*«В то время как; до того времени как.»* в БТС, *«В течение того времени как.»* в ТСО, *«Союз с фоновым значением (’в то время как").»* в «Двух веках в двадцати словах»)
3. До того времени как. *Пока солнце не взойдёт, на траве лежит иней.* (*«В то время как; до того времени как.»* в БТС, *«Союз с предельным значением ("вплоть до того как’).»* в «Двух веках в двадцати словах»)
4. Употребляется при прощании, до свидания. *Ну, я пошел, пока!* (*«Приветствие при прощании, до свидания.!»* в ТСО, *«Элемент формулы прощания.»* и *«Этикетное слово — до свидания.»* в «Двух веках в двадцати словах»)

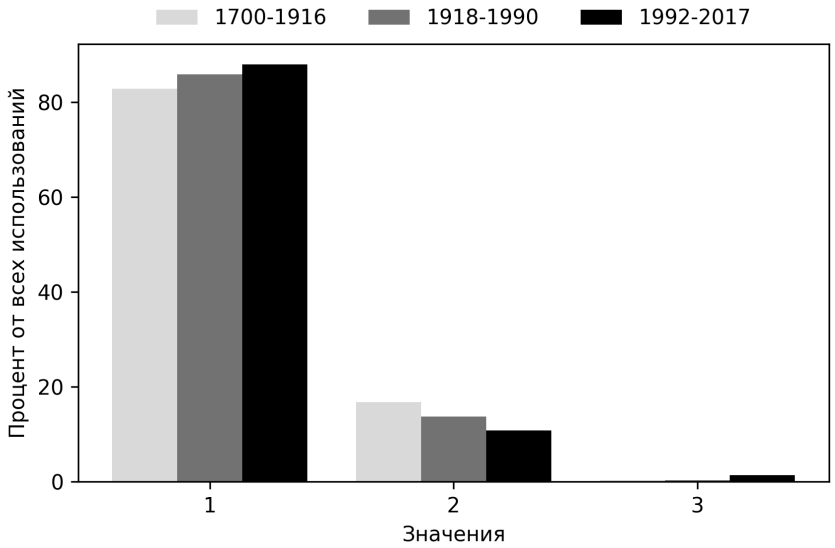


Рис. 8. Изменение значений слова *пока*

Значения для визуализации слова *пока* (Параметры: eps=0.25, min_samples=5).

1. В настоящее время, до тех пор.

2. Употребляется при обозначении времени, в течение которого совершается действие.
3. Употребляется при прощании с кем-л.

Анализ значений слова *пока*

Первое, второе и третье определения корректно сформулированы.

- ’В настоящее время, до тех пор.’ имеет общий смысловой элемент с ’В течение некоторого времени; до сих пор ещё; впредь до чего-л.’, а именно семы *’время’*, *’до сих пор’*, *’в течение’*.
- ’Употребляется при обозначении времени, в течение которого совершается действие.’ полностью соответствует ’В то время как.’, так как включает те же семы *’время’*, *’совершение действия’*.
- ’Употребляется при прощании с кем-л.’ полностью соответствует ’Употребляется при прощании, до свидания.’, так как включает те же семы *’прощание’*, *’до свидания’*.

Обобщённые значения, не найденные в визуализации:

- ’До того времени как.’ отсутствует среди предложенных моделью значений. Это значение близко к ’В то время как’, но с акцентом на предельность времени, что могло привести к отсутствию этого значения в визуализации.

Таким образом, для лексемы *пока* представлены 3 корректных определения.

Перейдем к частотности значений.

В книге, данные которой представлены на Рисунке 9, сообщается, что изначальным и всегда преобладающим было использование слова в качестве союза, после чего в XIX веке появилось использование как наречие, а затем в советский период – как этикетное слово. Данные из визуализации предложенного подхода поддерживают появление значения ’Употребляется при прощании с кем-л.’ поздно – несколько процентов для постсоветского перио-

да, однако данные для наречия и союза не совпадают. Можно предположить, что модели сложно различать эти значения из-за их схожести. Например, для *«Когда мы забирали щенка, нас предупредили, что ей категорически нельзя наверх забираться, пока у нее слабые лапы.»* было сгенерировано ’В настоящее время, до тех пор.’, что относит его к наречию, но из примера видно, что *пока* связывает части предложения и является союзом.

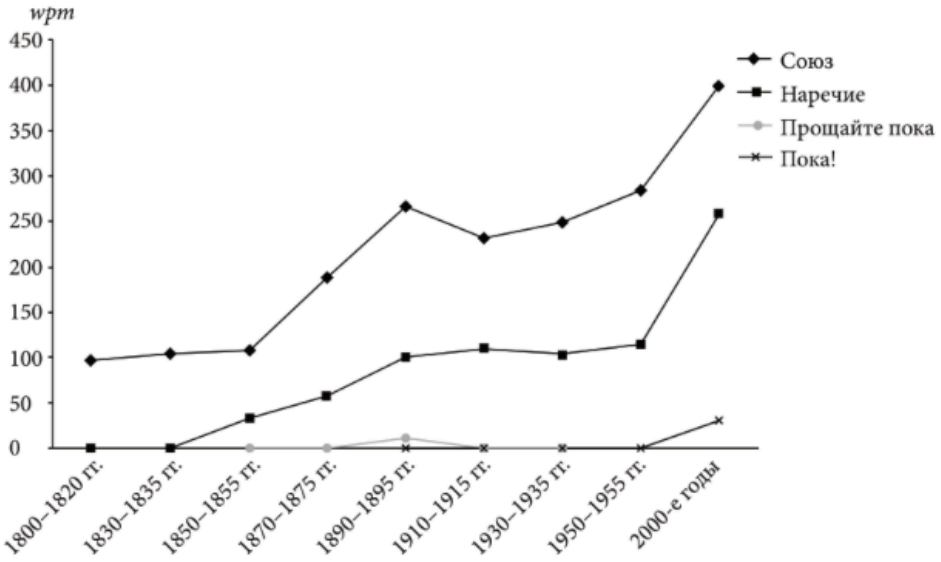


Рис. 9. Значения слова *пока* согласно (Данова [et al.], 2018)

Таким образом, предложенный подход лишь по большей части не отражает значения, в которых использовалось слово *пока*, так как из 3 выделенных значений, хоть и правильно сформулированных, статистика использования не согласуется с данными книги.

Анализ значений слова *пакет*

1. Упакованный в бумажную или иную обёртку какой-л. предмет (предметы); свёрток. *Пакет сахарного песка.* (*«Упакованный в бумажную или иную обёртку какой-л. предмет (предметы); свёрток.»* в БТС, *«Бумажный сверток, упаковка с чем-н.»* в ТСРЯ, *«Предмет, который завёрнут в бумажную или другую упаковку.»* в ТСД, *«Упаковка, сверток»* в «Двух веках в двадцати словах»)

2. Конверт с письмом официально-делового содержания. *Он вскрыл пакет и углубился в чтение.* (*«Конверт с письмом официально-делового содержания.»* в БТС, *«Конверт с письмом официального назначения.»* в ТСРЯ, *«Конверт с письмом официально-делового содержания.»* в ТСД, *«Письмо, конверт, почтовое отправление»* в «Двух веках в двадцати словах»)
3. Бумажный или полиэтиленовый мешок для упаковки каких-л. предметов, продуктов и т.п. *Продажа овощей в пакетах.* (*«Бумажный кулёк для упаковки каких-л. предметов, продуктов и т.п.»* в БТС, *«Бумажный мешок для продуктов, кулек.»* в ТСРЯ, *«Бумажный или полиэтиленовый кулёк с ручками или без для упаковки каких-либо предметов, продуктов и т. п.»* в ТСД, *«Ёмкость, тара»* в «Двух веках в двадцати словах»)
4. Комплект документов, официальных бумаг. *Пакет требований забастовщиков.* (*«Комплект документов, официальных бумаг.»* в БТС, *«В нек-рых сочетаниях: комплект документов, официальных бумаг.»* в ТСРЯ, *«Комплект документов или официальных бумаг.»* в ТСД)
5. Стопка ящиков или одинаковых деталей, строительных материалов и т.п., уложенных на специальный поддон для погрузки, перевозки и т.п. *Пакет труб.* (*«Стопка ящиков или одинаковых деталей, строительных материалов и т.п., уложенных на специальный поддон для погрузки, перевозки и т.п.»* в БТС, *«Стопка грузов, уложенная на поддон (спец.).»* в ТСРЯ, *«Комплект одинаковых деталей, строительных материалов и т. п.»* в ТСД)
6. Некоторое число акций какого-либо предприятия или компании, которым владеет человек или какая-либо организация, предприятие. *Контрольный пакет.* (*«Пакетом акций является некоторое число акций какого-либо предприятия или компании, которым владеет*

человек или какая-либо организация, предприятие.» в ТСД, *«Пакет акций»* в «Двух веках в двадцати словах»)

7. Совокупность информации, собранной для разовой передачи по компьютерной сети. *Пакет данных.* (*«Совокупность информации, собранной для разовой передачи по компьютерной сети.»* в БТС)
8. Набор взаимосвязанных элементов, объединённых общей целью. *Пакет льгот и скидок.* (*«Наборы» ('нематериальная совокупность')* в «Двух веках в двадцати словах»)

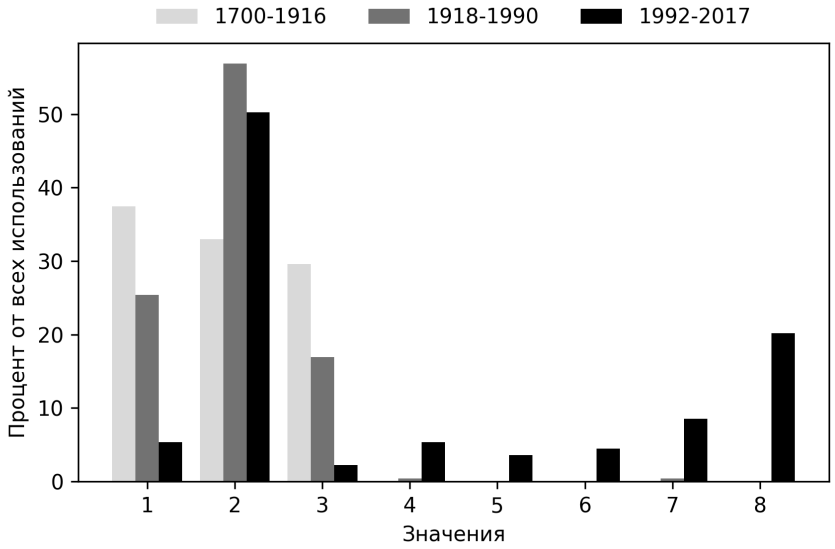


Рис. 10. Изменение значений слова *пакет*

Значения для визуализации слова *пакет* (Параметры: eps=0.11, min_samples=8).

1. Письмо, посылка и т. п. в таком виде.
2. Бумажный или матерчатый мешочек с чем-либо для хранения, перевозки и т. п.
3. Письмо, посылка и т. п., запечатанные в такой конверт.
4. Совокупность каких-либо однородных, связанных между собой предметов, явлений и т. п.
5. Совокупность программных средств, объединенных по какому-либо признаку.

6. Часть чего-либо, принадлежащая кому-либо на определенных условиях.
7. Совокупность каких-либо однородных предметов, документов и т. п.
8. Совокупность акций какого-либо акционерного общества.

Анализ значений слова *пакет*

Все определения, кроме шестого, корректно сформулированы.

- ’Письмо, посылка и т. п. в таком виде.’, а также ’Письмо, посылка и т. п., запечатанные в такой конверт.’ имеет общий смысловой элемент с ’Конверт с письмом официально-делового содержания.’, а именно семы *’письмо’*, *’посылка’*, *’конверт’*.
- ’Бумажный или матерчатый мешочек с чем-либо для хранения, перевозки и т. п.’ соответствует ’Бумажный или полиэтиленовый мешок для упаковки каких-л. предметов, продуктов и т.п.’, общие семы *’мешок’*, *’бумажный’*.
- ’Совокупность программных средств, объединенных по какому-либо признаку.’ частично соответствует ’Набор взаимосвязанных элементов, объединённых общей целью.’, так как включает те же семы *’совокупность’*, *’объединенных объектов’*, но является более узким, так как касается только программных средств. Среди примеров, которые были выделены алгоритмом, находятся такие, как *Для обработки же растровых изображений и конкретно цифровых фотографий у компании ”Corel” существует пакет Corel Paint Shop Pro Photo.*, где значение слова действительно может быть описано как ’Совокупность программных средств, объединенных по какому-либо признаку.’, поэтому мы будем считать это определение корректным.

- ’Совокупность акций какого-либо акционерного общества.’ полностью соответствует ’Некоторое число акций какого-либо предприятия или компании, которым владеет человек или какая-либо организация, предприятие.’, так как включает те же семы *’совокупность’*, *’акции’*.
- ’Совокупность каких-либо однородных, связанных между собой предметов, явлений и т. п.’, , а также ’Совокупность каких-либо однородных предметов, документов и т. п.’ соответствует ’Набор взаимосвязанных элементов, объединённых общей целью.’.
- ’Часть чего-либо, принадлежащая кому-либо на определенных условиях.’ не имеет схожих определений среди обобщенных и является некорректным. Анализ примеров, для которых алгоритм дал такое определение, показывает, что большинство примеров связано с акциями, например, «Но контрольный пакет акций был размыт.». В данном случае логичным является определение, акцентирующее *’совокупность’*.

Обобщённые значения, не найденные в визуализации:

- ’Упакованный в бумажную или иную обёртку какой-л. предмет (предметы); свёрток.’ отсутствует среди предложенных моделью значений.
- ’Стопка ящиков или одинаковых деталей, строительных материалов и т.п., уложенных на специальный поддон для погрузки, перевозки и т.п.’ также отсутствует в визуализации. Это значение акцентируется на физической стопке предметов (ящиков, деталей) на поддоне, что могло быть причиной отсутствия в визуализации.

Таким образом, для лексемы *пакет* представлено 7 корректных определений. Среди ошибок присутствуют:

- Некорректность определения: 1

Перейдем к частотности значений.

Основные изменения в значениях слова *пакет*, указанные в книге «Два века в двадцати словах», расположены на Рисунке 11. В досоветский период значения, обозначающие ’почтовое отправление’ и ’упаковка, свёрток’ уступили место ’мешку, в том числе из полиэтилена’, ’набору’ и ’картонной упаковке’. Информация из визуализации предложенного подхода частично соответствует этим данным. ’Письмо, посылка и т. п. в таком виде.’ и ’Письмо, посылка и т. п., запечатанные в такой конверт.’ вместе преобладают в досоветский период, уступая место значению ’Бумажный или матерчатый мешочек с чем-либо для хранения, перевозки и т. п.’ в советское время. Кроме того, согласно алгоритму, в постсоветское время появляются такие значения, как ’Совокупность каких-либо однородных, связанных между собой предметов, явлений и т. п.’, ’Совокупность акций какого-либо акционерного общества.’, что соответствует информации из книги.

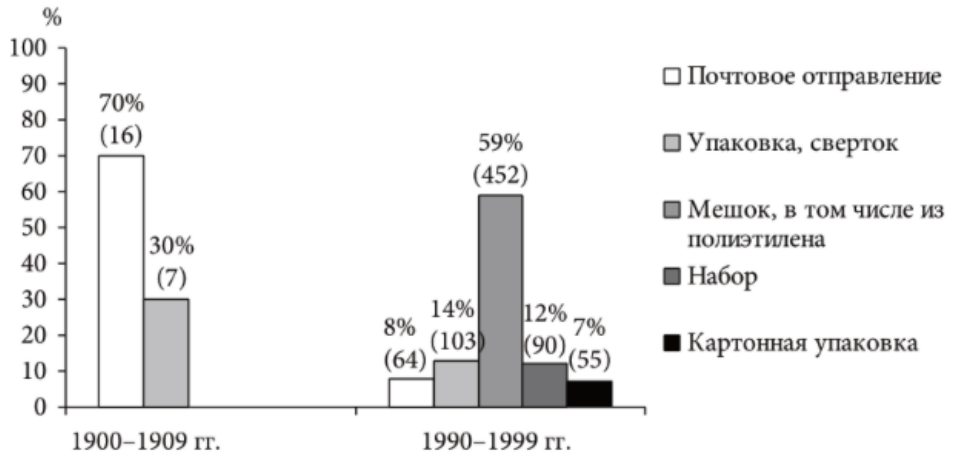


Рис. 11. Значения слова *пакет* для 1900-1999 согласно (Данова [et al.], 2018)

Таким образом, предложенный подход в целом отражает значения, в которых использовалось слово *пакет*, согласуясь с данными из толкового словаря и историческим исследованием.

3.8. Результаты качественного анализа

В результате обобщения словарных дефиниций было составлено 119 значений для 20 слов.

Всего в результате работы предложенного подхода было получено 92 определения для 20 слов.

Таким образом, без учёта некорректных определений было успешно выявлено 73.1% значений.

Таблица 12. Типы определения и их количество

Тип определения	**Количество**	**Процент**
Корректные	57	61.95%
Близкие	10	10.87%
Некорректные	5	5.43%
Недостаточно конкретизированные	5	5.43%
Избыточность или чрезмерное использование общих фраз	4	4.35%
Близкое значение, а также избыточность или чрезмерное использование общих фраз	1	1.09%
Избыточно конкретизированные	1	1.09%
Самореференция	0	0.00%
Противоположное значение	0	0.00%
Неправильная часть речи	0	0.00%

Как видно, из результатов большинство определений являются корректными без каких-либо ошибок или недочётов (61.95%).

Кроме того, проблема с самореференцией была успешно решена.

Из частых проблем можно выделить:

- Близкое значение Примером можно привести определение ’Насекомое, похожее на червя, а также его личинка.’ для слова *червяк*, где допущена ошибка, так как червяк не может быть взрослым насекомым. Возможно, часть таких ошибок связана с относительно небольшим размером модели, что не позволяет ей иметь уверенные знания об окружающем мире.
- Некорректные Примером является ’Употребляется для присоединения предложений или отдельных членов предложений, усиливающих или уточняющих высказанную мысль.’ для слова *пожалуй*. Как видно, определение слова *пожалуй* не соответствует его правильному значению, что также может быть связано с ограничениями модели.
- Избыточность или чрезмерное использование общих фраз В нашем случае эта проблема проявляется в повторении слов в определении. Например, для слова *свалка* в сгенерированном определении ’Беспорядочная, беспорядочная схватка.’ повторяется слово *беспорядочная*. На наш взгляд, именно такая ошибка может быть связана с обилием синонимических рядов в определениях обучающего датасета на основе «Малого академического словаря», что является одним из способов описать значение слова в лексикологии.
- Недостаточное конкретизирование Например, ’Ласковое обращение к женщине.’ для слова *мама* близко к ’Обращение ребёнка к своей матери.’, но имеет более широкий смысл, включающий всех женщин, а не только матерей или нянь, что делает его недостаточно специфичным.

Кроме того, одной из выявленных проблем является недостаток контекста для разграничения значения. Возьмём как пример слово *пионер*.

Для него новым значением является ’Член добровольной самодеятельной детской организации.’, появившееся в советское время. В визуализации

указано около 20% использований слова в таком значении в досоветский период, что объясняется двусмысленностью части примеров, например, *Пионеры слушают это и восхищаются.*, где необходим дополнительный контекст для установления значения.

К сожалению, для 2 слов из списка представляется невозможным полноценно проанализировать статистическую информацию. Этими словами являются *публика* и *сволочь*.

Так, для слова *публика* в книге «Два века в двадцати словах» не даётся диаграмм частотности и точных для слова *публика*. Говорится лишь о преобладании значения ’аудитория’ и о его оттенках, которые не удается полноценно сравнить из-за того, что алгоритм предложил довольно общие значения.

Для слова *сволочь* из 4 значений, выявленных при обобщении дефиниций, предложенным подходом были выявлены только следующие два значения:

1. Употребляется как бранное слово.
2. О подлом, гнусном человеке.

К сожалению, оба выделенных значения подпадают под значение ’Индивидуальное оскорбление.’ в книге «Два века в двадцати словах», поэтому анализ изменений значения сделать не представилось возможным.

Кроме того, затруднителен анализ для слова *кануть*. Для него подтверждается информация о том, что с 1900 года ’исчезнуть, сгинуть, пропасть’ является основным значением, однако книга не предоставляет визуализаций частоты использования значений слова по периодам.

Среди большинства оставшихся слов визуализации в какой-то совпадают с данными из книги «Два века в двадцати словах». Исключением является рассмотренное ранее слово *пока*, где результаты визуализации противоречат данным исследования.

Несмотря на это, можно сказать, что основные изменения значения были уверенно выявлены в 12 словах – большинстве.

Одной из самых качественных визуализаций была сделана для слова *пакет*. Так, в нём было выявлено 7 корректных определений, 4 из которых встречаются только в постсоветский период:

- Совокупность каких-либо однородных, связанных между собой предметов, явлений и т. п.
- Совокупность программных средств, объединенных по какому- либо признаку.
- Совокупность каких-либо однородных предметов, документов и т. п.
- Совокупность акций какого-либо акционерного общества.

2 определения, связанные со значением ’письма’ теряют в популярности в советский и постсоветский период, а значение ’бумажного или матёрчатого мешочка’ растёт в использовании.

Все вышеперечисленные результаты согласуются с данными из исследования истории данных слов.

Кроме того, в 4 словах изменения были выявлены и частично согласуются.

Таким образом, представляется возможным утверждать, что моделирование определений даёт интерпретируемые описания значений слов, которые могут быть успешно использованы для выявления семантических изменений.

Выводы

Таким образом, была дообучена большая языковая модель FRED-T5-1.7B для задачи генерации определений с помощью датасета на основе «Малого академического словаря». Проведенные эксперименты показали, что модель способна генерировать семантически близкие, но не всегда идентичные определения. Несмотря на сравнительно низкие значения метрик BLEU и ROUGE-L, отражающих формальное сходство сгенерированных опреде-

лений с эталонными, модель демонстрирует высокие результаты по метрике BERT-F1, учитывающей семантическую близость текстов. Это говорит о том, что модель способна продуцировать определения, имеющие схожий смысл с эталонными, хоть и зачастую сформулированные иными словами.

Применение модели для анализа семантических сдвигов на материале соревнования Rushifteval показало, что решение на основе FRED-T5-1.7B вместе с дообучением векторизатора способно добиться высокого результата в лидерборде. Таким образом, данная модель демонстрирует хорошие результаты в задаче выявления семантических изменений.

Разработанная визуализация на основе кластеризации векторных представлений определений позволяет наглядно представить семантические изменения слов во времени, что может быть полезно для лингвистических и исторических исследований.

Кроме того, была произведена качественная оценка работы модели на примере произведённых с её помощью визуализаций, в ходе которого было определено, что большинство из сгенерированных определений являлись корректными, а статистическая информация по изменению использования отдельных значений во времени преимущественно совпадала с результатами экспертного исследования истории слов.

Заключение

Таким образом, в ходе выполнения выпускной квалификационной работы была обучена генеративная языковая модель на основе материала словаря МАС для задачи генерации определений слов на основе их контекста использования. Кроме того, были выполнены следующие задачи:

- разработан алгоритм автоматического определения семантических сдвигов на основе векторного представления сгенерированных определений;
- проведён анализ метрик и качества обученной языковой модели и сравнение её с существующими аналогами;
- был разрабон алгоритм визуализации результатов;
- проведён качественный анализ результатов предложенного подхода;

Модель показала высокие результаты метрики сходства BERTScore для тестовой выборки, а также успешно показала себя на тестовом материале Rushifteval, имея результаты сопоставимые с лидирующими решениями. При качественном анализе результатов предложенный подход показал высокие результаты, выявив большинство значений, а также верно составив визуализацию изменения использования значений для большинства слов. В процессе выполнения настоящей работы было доказано, что моделирование определений может быть успешно применено для задачи детектирования семантических изменений, а также позволяет достичь высокого уровня интерпретируемости.

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 (Giulianelli [et al.], 2023). 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 (Gardner [et al.], 2022).

Перспективами развития настоящего исследования является:

- Использование более одного словаря в качестве материала для обучения модели,
- Использование генеративных моделей большего размера, чем используется в работе.

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. (URL: https://github.com/tatarinovst2/work-definition-modeling)

Список литературы

1. *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.
2. *Данова М. К.*, *Добрушина Н. Р.*, *Опачанова А. С.* [et al.]. Два века в двадцати словах. — Москва : Издательский дом Высшей школы экономики, 2018. — 455 с.
3. *Tahmasebi N.*, *Borin L.*, *Jatowt A.* Survey of computational approaches to lexical semantic change detection // Computational approaches to semantic change. — Berlin : Language Science Press, 2021. — С. 1—91.
4. *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.
5. *Rodina J.*, *Trofimova Y.*, *Kutuzov A.*, *Artemova E.* ELMo and BERT in semantic change detection for Russian. — 2020.
6. *Giulianelli M.*, *Luden I.*, *Fernández R.*, *Kutuzov A.* Interpretable Word Sense Representations via Definition Generation: The Case of Semantic Change Analysis. — 2023.
7. *Gardner N.*, *Khan H.*, *Hung C.-C.* Definition modeling: literature review and dataset analysis // Applied Computing and Intelligence. — 2022. — Т. 2. — С. 83—98.
8. Sketch engine. — 26.05.2024. — URL: https://www.sketchengine.eu/guide/word-sense-induction/.

9. *Periti F.*, *Cassotti P.*, *Dubossarsky H.*, *Tahmasebi N.* Analyzing Semantic Change through Lexical Replacements. — 2024.

10. *Майсак Т. А.* Грамматикализация. — 2016 ; — Accessed: 2024-05-21. *Большая российская энциклопедия. Электронная версия*.

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

12. *Jatnika D.*, *Bijaksana M.*, *Ardiyanti A.* Word2Vec Model Analysis for Semantic Similarities in English Words // Procedia Computer Science. — 2019. — Янв. — Т. 157. — С. 160—167.

13. *Mikolov T.*, *Chen K.*, *Corrado G.*, *Dean J.* Efficient Estimation of Word Representations in Vector Space. — 2013.

14. *Ester M.*, *Kriegel H.-P.*, *Sander J.*, *Xu X.* A density-based algorithm for discovering clusters in large spatial databases with noise // Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. — Portland, Oregon : AAAI Press, 1996. — С. 226—231.

15. *Kutuzov A.*, *Fomin V.*, *Mikhailov V.*, *Rodina J.* ShiftRy: Web service for diachronic analysis of Russian news //. — 01.2020. — С. 500—516.

16. *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.

17. *Kutuzov A.* Distributional Word Embeddings in Modeling Diachronic Semantic Change : Doctoral Thesis / Kutuzov Andrey. — University of Oslo, 2020. — Accessed: 2020-11-16T12:34:15Z.

18. *Kutuzov A.*, *Pivovarova L.* RuShiftEval: a shared task on semantic shift detection for Russian // Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue. — 2021. — С. 533—545.

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

20. *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. — T. abs/1911.02116.

21. *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. — C. 16—30.

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

23. *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. — C. 266—271.

24. *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. — C. 2499—2509.

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

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

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

28. *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.

29. *Ахапкин Д.* [et al.]. Толковый словарь русского языка : Ок. 2000 слова. ст., свыше 12000 значений. — Москва : Астрель [и др.], 2003. — С. 989. — ГУП ИПК Ульян. Дом печати.

30. *Ушаков Д.* Толковый словарь русского языка. Т. 4. С - Ящурный. — Москва : Государственное издательство иностранных и национальных словарей, 1940.

31. *Zmitrovich D.* [et al.]. A Family of Pretrained Transformer Language Models for Russian. — 2023.

32. *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.

33. *Евгеньева А. П.* Словарь русского языка: В 4-х т. — Москва : Русский язык, 1981-1984. — В 4-х томах.

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

35. paraphrase-multilingual-mpnet-base-v2. — URL: https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2 (дата обр. 19.04.2024).

36. encodechka. — URL: https://github.com/avidale/encodechka (дата обр. 19.04.2024).

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

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

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