ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет

«Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа

«Компьютерная лингвистика»

Подрядчикова Мария Владимировна

АВТОМАТИЧЕСКИЙ АНАЛИЗ НАРРАТИВНОГО ПОВЕДЕНИЯ НА МАТЕРИАЛЕ ДНЕВНИКОВЫХ ЗАПИСЕЙ

Выпускная квалификационная работа студента 2 курса магистратуры группы МКЛ181

Академический руководитель

образовательной программы
канд. филологических наук, доц.
А.А. Бонч-Осмоловская

_____

« » _____ 2020 г.

Научный руководитель

канд. филологических наук, доц.
А.А. Бонч-Осмоловская

_____

Москва 2020

Оглавление

Вступление

Развитие методов автоматической обработки текста и появление больших оцифрованных корпусов способствовало распространению исследований, посвященных “дальнему чтению” (distant reading) — подходу, который использует методы автоматической обработки информации при изучении литературы. В статье 2000 года, в которой Франко Моретти (Moretti 2000) вводит этот термин, он исследует общие закономерности, характерные для развития литературы разных стран. С тех пор методы дальнего чтения использовались при анализе текстов различных форм и жанров: от корпусов драмы (Fischer et al. 2016) до политических тредов на Reddit (Aurnhammer et al. 2019).

В отличие от традиционного “медленного чтения” (close reading), которое опирается на глубокий анализ достаточно ограниченного канона текстов, дальнее чтение позволяет исследовать большие корпуса произведений и анализировать объекты, как меньшие, чем произведение, — например, тропы, так и большие — например, жанры (Moretti, 2000). У подхода есть и свои ограничения: при отсутствии обширного корпуса текстов анализировать глобальные закономерности крайне затруднительно. Как следствие, появление такого корпуса для определенного жанра текстов создаёт новые возможности для его исследования.

Наше исследование стало возможным благодаря появлению в 2014 году текстового корпуса личных дневников “Прожито”. На сентябрь 2019 года библиография проекта превышала 4000 дневников различных авторов. Значительную часть корпуса составляют расшифровки рукописей, что позволяет расширить материал для изучения дневниковых записей. Если раньше преимущественно изучались опубликованные дневники известных людей (в особенности писателей), то среди авторов дневников "Прожито" есть школьники, преподаватели, врачи, агрономы, — те, чьи дневники до этого редко были доступны исследователям.

Цель нашей работы — научиться автоматически выявлять общие закономерности в дневниках разных авторов. Поскольку тематика дневников разнообразна, мы будем опираться на признаки, независимые от жанра дневника и вида деятельности его автора. Для этого мы разработаем классификацию дневниковых записей, основанную на интенции автора — его потребности выразить в записи информацию определенного типа: например, описание отдельного эпизода из жизни или выражение чувств. Особенности проявления этих интенций в корпусе дневниковых записей одного автора мы будем называть его нарративным поведением.

Новизна работы заключается в создании собственной классификации дневниковых записей и исследовании возможностей её применения для дальнего чтения дневников. Мы ожидаем, что изучение нарративного поведения позволит в перспективе выявить новые закономерности в дневниках разных авторов, разных жанров и разных периодов истории.

Работа строится следующим образом. В первой главе работы будет дан научный контекст исследования: мы рассмотрим развитие подходов к формальному представлению текста и современные исследования, посвященные этой теме, а также отдельно будет представлен обзор теоретических работ, посвященных специфике дневникового нарратива.

Во второй главе мы предложим собственную классификацию дневниковых записей. Мы формализуем особенности различных дневниковых записей и выделим на этой основе четыре категории, охватывающие большую часть записей корпуса независимо от личности автора записи, его рода деятельности или времени написания дневника. Там же будет представлено обсуждение процедуры разметки по выделенным категориям и случаи несовпадения мнений разметчиков при разметке корпуса.

В третьей главе мы применим методы автоматической классификации текста к размеченным нами записям и выберем модель, наиболее точно определяющую категорию записи.

В четвертой, последней, главе мы сделаем первые шаги в применении полученной модели к дальнему чтению корпуса дневников “Прожито”. Мы рассмотрим связь особенностей нарративного поведения с возрастом авторов дневников: как на конкретных примерах, так и на более общей выборке из корпуса.

Таким образом, в нашей работе будет реализован подход к решению задачи автоматического выявления общих закономерностей в различных корпусах дневников. Мы предложим универсальную классификацию дневниковых записей, которая позволит находить схожие интенции в дневниках разных жанров. На её основе мы создадим алгоритм автоматического определения интенции дневниковой записи, который затем применим для дальнего чтения корпуса дневников.

1. Обзор литературы

Итак, целью исследования является автоматическое извлечение общих закономерностей в дневниковых нарративах. Поэтому научный контекст исследования можно рассматривать в трёх основных измерениях:

— история и развитие формализации нарративной структуры текста

— особенности дневника как жанра и подходы к классификации дневников

— автоматический анализ нарративной структуры текста

Первое из них задаёт парадигму исследования, второе специализирует его объект, а третье — метод.

1.1. Исследования формализации нарратива

Формализация нарративной структуры текста позволяет в том числе автоматически анализировать различные документы: сравнивать их между собой, выделять в них сходства и различия, искать особенности, характерные для разных жанров. Для того, чтобы исследовать общие закономерности в корпусе дневников, мы тоже формализуем ряд особенностей нарративного поведения авторов. В этом разделе мы рассмотрим то, как разные авторы подходили к задаче формализации нарратива.

Одним из основополагающих исследований нарратива со структурно-типологической точки зрения является “Морфология сказки” В. Проппа (Пропп 1928). В своей работе Пропп критикует существующие на тот момент классификации сказок, часто основанные на некоторых ярких особенностях сказки (характеристике героев, теме завязки и т.п.) и не учитывающие структурные особенности сказочного повествования. Он предлагает разработать классификацию, основанную на структурных признаках, и выделяет в качестве таких признаков функции действующих лиц — определенные действия, которые часто повторяются в различных сказках различными персонажами.

Во второй половине XX века появляются более универсальные формализмы, позволяющие отразить структуру разнообразных историй. К ним можно отнести, например, работы В. Ленерт о сюжетных единицах (plot units) (Lehnert 1981) и В. Лабова и Д. Валетски о структуре нарратива (Labov, Waletzky 1967). В отличие от работы Проппа, в подобных исследования рассматриваются единицы структуры нарратива, которые можно выделить в текстах **различных** жанров и стилей. При этом всё ещё важным является наличие в тексте описания каких-либо событий/изменений.

В исследовании Ленерт основной единицей структуры является “affect state” — состояние, в котором может пребывать персонаж: ментального (предположительно нейтрального), положительного или отрицательного. Ленерт описывает возможные последовательности этих действий и показывает различные причинно-следственные и временные зависимости в нарративе. С этой точки зрения анализируются как короткие истории из нескольких предложений, так и более длинные нарративы — например, рассказ О. Генри “Дары волхвов” (Lehnert 1981).

Другой подход к описанию структуры нарратива предлагают Лабов и Валетски (Labov, Waletzky, 1967). По их классификации, нарратив состоит из следующих частей: Orientation, Complication, Evaluation, Resolution и Coda. Каждая из них имеет собственные функции: например, Orientation представляет действующих лиц и время и место действия, а Complication описывает отдельное событие нарратива. Данные части достаточно универсальны и могут быть выделены в текстах различных жанров.

Формализмы, подобные формализму Лабова и Валетски, нередко используются при описании нарратива и в современных работах (Li et al. 2018) (Tangherlini 2018) (Al Schboul 2018). К примеру, в статье “Annotating High-Level Structures of Short Stories and Personal Anecdotes” (Li et al. 2018) используется более подробная схему разметки из 10 категорий и рассматривается согласованность разметки одинаковых историй разными разметчиками. Авторы замечают, что различия между некоторыми категориями, например, оценкой

(Evaluation) и последствием (Aftermath), не всегда очевидны для разметчиков. При объединении схожих категорий согласованность заметно увеличивается. По-видимому, подробные схемы разметки не всегда удаётся формализовать достаточно однозначно.

В этом разделе мы рассмотрели некоторые формализмы нарративной структуры текста. Как правило, они применяются к коротким текстам с чётко выраженной последовательностью событий: например, сказкам или небольшим рассказам. В следующем разделе мы рассмотрим особенности дневникового нарратива для того, чтобы понять, каким образом следует формализовать тексты этого жанра.

1.2. Особенности дневника как жанра

На данный момент существует малое количество исследований, посвященных анализу нарративной структуры дневников, и, соответственно, отсутствуют формализмы, описывающих именно дневники. Для того, чтобы лучше представлять себе возможность формальной классификации дневниковых записей, рассмотрим само определение дневника, особенности этого жанра и существующие классификации дневников.

Согласно определению из Литературной энциклопедии терминов и понятий (Жожикашвили 2003), дневник – это периодически пополняемый текст, состоящий из фрагментов с указанной датой для каждой записи. Данное определение не отражает многие из значимых функций личного дневника, выделенных филологами.

М.Ю. Михеев, говоря о дневнике, в первую очередь отмечает, что дневники относятся к так называемым эго-текстам — текстам, которые “имеют в целом автобиографическую направленность, будучи обращены на мир из субъективной точки зрения, уникального *здесь и теперь* своего автора и центрированы вокруг субъекта” (Михеев 2006).

А.А. Зализняк считает главным признаком дневника то, что автор в

дневнике является адресатом, и, кроме того, существует потенциальный второй, косвенный адресат. Другие признаки включают в себя неотделимость автора от повествователя, нефикциональность текста дневника, отсутствие единого авторского замысла, а также то, что дневник — это в первую очередь текст о себе и текущем моменте (Зализняк 2010).

При классификации дневников их часто рассматривают с точки зрения деятельности ведущего их человека. Разграничивают “профессиональные” дневники (журналистов, писателей и т.п.) и “непрофессиональные” , выделяют дневники путешественников, врачей и деятелей искусства (Михеев 2006). Любопытно, что в 2006 году Михеев называет дневники писателей “самой распространенной разновидностью” дневников (Михеев 2006): до появления корпуса значительная часть дневников “обывательских”, по-видимому, была вне исследовательского поля зрения.

В одной из значительных работ, представляющих сразу несколько классификаций дневника, “Русском литературном дневнике XIX века” О.Г. Егорова (Егоров 2011), анализирует как раз дневники русских писателей. Исследование является классическим примером “медленного чтения”: изучив несколько десятков дневников, автор выделяет типы, исходя из различных особенностей текстов. Дневники рассматриваются с точки зрения психологических особенностей (экстравертивный, интровертивный, переходящий и осциллирующий типы), жанрового содержания (семейно-бытовой, путевой, общественно-политический, служебный), а также стилевой формы (информативно-повествовательный и аналитический стили, эстетически нагруженное слово и смешанная форма).

Все найденные нами классификации объединяет один признак — основной единицей в них является сам дневник, то есть полный корпус дневниковых записей одного автора, даже в том случае, когда исследователь называет форму этого корпуса “смешанной”. Таким образом, подобного рода классификация будет недостаточна для нашего исследования. Во-первых, подобные классификации преимущественно описывают "профессиональные" дневники и не

отражают всё разнообразие корпуса, с которым мы работаем. Во-вторых, для того, чтобы изучать нарративное поведение автора, мы должны исследовать в том числе внутреннюю структуру дневника, которая может быть неоднородной. В этом случае будет недостаточно сказать, что мы изучаем "дневник такого-то типа": необходимо будет выделить разные категории для разных записей дневника. Следовательно, нам придётся создать собственную классификацию, опираясь на уже существующие (наиболее полезной и, что важно, не зависящей от личности автора кажется классификация стилевой формы О.Г. Егорова).

В следующем разделе этой главы мы рассмотрим то, как формализмы используются при автоматическом анализе текста. Мы опишем современные исследования, основанные на материале, частично напоминающем дневниковые записи: на художественной литературе, записях блогов и постах на сайте Reddit.

1.3. Автоматический анализ нарративной структуры текста

Сложные формализмы не всегда удаётся эффективно использовать при разметке текстов. Как можно было видеть в первом разделе главы на примере работы “Annotating High-Level Structures of Short Stories and Personal Anecdotes” (Li et al., 2018), при уменьшении количества категорий возрастает согласованность оценок разметчиков и удобство формализации нарратива. При автоматическом анализе нарративной структуры часто используются упрощенные способы формализации или анализируются только часть информации, изложенной в тексте. Рассмотрим то, как исследователи выявляют разные формальные особенности структуры текста при анализе разных жанров.

Один из простых методов формализации предлагает выделять в тексте только конкретные сюжетные единицы. На нём основана, например, работа “Literary Event Detection” (Sims et al., 2019), в которой авторы выясняют, что менее престижные произведения отличаются большей плотностью событий. Авторы решают задачу бинарной классификации, определяя, является ли токен словом, означающим начало события (“event trigger”) или нет. При разметке была показана высокая согласованность оценок: каппа Коэна (Cohen 1960) = 0,813. Для

автоматической бинарной классификации токенов использовались такие признаки, как эмбеддинги, обученные на корпусе художественной литературы, а также, для некоторых моделей, частеречные теги.

В работе “Modelling Protagonist Goals and Desires in First-Person Narrative” (Rahimtoroghi et al. 2017) авторы решают конкретную задачу: определяют, исполнилось или нет желания автора записи в блоге, в которой это желание было описано. Запись делится на четыре части: “выражение желания”, предшествующий и последующий контексты и “подтверждение” (тому, что желание осуществилось или не осуществилось). При разметке того, осуществилось желание или нет, для 66% записей было достигнуто полное согласие оценок трёх разметчиков. При автоматической классификации в базовом решении использовалось представление текста в виде мешка слов (bag of words), а в последующих — предобученные эмбеддинги. Также были использованы дополнительные семантические признаки, отражавшие тональность текста и некоторые дискурсивные особенности. Их использование помогло повысить качество классификации.

Задача классификации, поставленная в работе “Using Functional Schemas to Understand Social Media Narratives” (Yan et al. 2019), кажется ближе всего к нашей задаче классификации дневниковых записей. Классами в ней являются разделы сайта Reddit, записи в которых схожи тематически (каждый из них посвящен экологии), но часто отличаются структурой нарратива. В данной работе авторы не предлагали собственный формализм, но использовали выделение определенных “функциональных схем” (например, “просьба о помощи” или “выражение собственного мнения”) для классификации постов наряду с лексическими признаками (предобученными эмбеддингами). Использование этих признаков повысило качество классификации.

В этой главе мы рассмотрели исследования в области формального представления и автоматического анализа нарратива, а также особенности жанра личного дневника и классификации дневников.

Мы проанализировали особенности различных формализмов и отметили,

что часто формализмы, которые используются для автоматического извлечения информации из текста, проще, чем классические формализмы.

Исследование особенностей и видов дневников, а также подходов к автоматическому анализу похожих видов нарратива, поможет нам в будущем решить задачи создания собственной классификации дневниковых записей и автоматического определения класса записи, которые мы рассмотрим в следующих главах.

2. Создание классификации дневниковых записей

2.1. Описание данных

Мы исследуем корпус личных дневников “Прожито”[1] (ссылка на корпус в приложении 2). Корпус состоит из 384587 записей 3464 авторов и охватывает период с XVII по XXI век. Некоторые из дневников уже были опубликованы ранее, некоторые являются оцифрованными волонтёрами центра рукописями.

Корпус представлен в виде нескольких связанных csv-таблиц. Для нашего исследования наиболее важной информацией была информация о записи (notes.csv) и об её авторе (authors.csv). Для чтения таблиц мы использовали библиотеку prozhito-tools[2].

Мы также будем использовать в работе информацию об авторе и записи: ID записи, текст записи, ID автора записи, дата написания записи, имя автора, год рождения автора (в ряде случаев).

Кроме того, мы использовали сайт центра “Прожито” для поиска дополнительной информации об интересующих нас авторах.

Источник всех записей, цитируемых в дальнейшем — данный корпус. Запись цитируется в том виде, в котором она была представлена в корпусе (включая оформление сокращений и дополнительных помет). В случае, если мы

1 Корпус центра изучения эго-документов “Прожито” — https://prozhito.org/
2 https://github.com/kilomeow/prozhito-tools

приводим цитату только частично, мы заменяем пропущенную часть отточием “[...]”. После каждой из записей в скобках указаны её автор и дата написания.

2.2. Разработка классификации дневниковых записей

Целью нашей работы является автоматическое выявление общих закономерностей в дневниках. Для этого мы разрабатываем формализм, который позволит нам описывать нарративное поведение авторов дневников.

Выше в обзоре литературы мы рассмотрели некоторые подходы к формализации нарративов, включающие классификацию нарративных единиц, а также обратились к классификациям дневников. Однако, на наш взгляд, ни одна из рассмотренных классификаций не позволяет в полной мере описать разнообразие особенностей жанра дневниковых записей. Например, рассмотрим классификации, основным объектом которых является фрагмент текста (событие или другая единица) (Labov, Waletzky 1967) (Lehnert 1981). Они предполагают, что мы анализируем определенный нарратив, как правило, насыщенный событиями. Это исключает возможность исследовать записи иной формы: например, полностью состоящие из размышлений или описания природы. Существующие классификации дневников, напротив, выделяют в качестве отдельной единицы целый корпус записей одного человека. Мы считаем, что не все дневники являются однородными, а нарративное поведение автора на протяжении ведения дневника может меняться. Следовательно, мы не можем выделить дневник в качестве основной единицы классификации.

Таким образом, мы приняли решение разработать собственную классификацию дневниковых записей. Ниже мы перечисляем список требований к ней.

а) Основная единица классификации — дневниковая запись.

Анализируется не весь дневник и не часть записи, а отдельная запись. С одной стороны, это позволит включить в классификацию дневники вне жанровой

системы дневников — в частности, более глубоко изучить записи дневников, которые ранее относили к достаточно общим категориям (например, “обывательские дневники”). С другой, это даст возможность включить в одну классификацию записи с разной внутренней структурой: как насыщенные событиями, так и записи-размышления и записи-описания.

б) Классификация должна быть универсальна.

Классификация не должна быть ограничена только тем материалом, который у нас уже есть. Предполагается, что любая новая запись может быть отнесена к одной из существующих категорий.

в) Критерии, по которым записи присваивается категория, прозрачны.

Признаки, характерные для категорий, должны быть чётко и однозначно сформулированы. Кроме того, классификация должна быть достаточно простой. Это предотвратит множество спорных случаев для схожих категорий при разметке, а также упростит потенциальную автоматическую классификацию записей.

2.2.1. Выделение значимых признаков и создание первого варианта классификации

Анализируя полученный нами корпус, мы приняли решение различать записи на основе их интенции — потребности выразить в записи информацию определенного типа, например, пересказать события или написать рабочий отчёт.

Мы выделили несколько признаков, отличающих тексты с разными интенциями друг от друга: предмет описания, насыщенность текста событиями и стиль речи.

а) Предмет описания

Во-первых, тексты дневниковых записей могут описывать как происходящие в физическом мире события, так и их восприятие автором и его размышления и чувства. Кроме того, некоторые из записей включают отрывки из

литературных произведений. Ниже в (1-3) приведены примеры таких записей

*1) “**Пятница.** Ночью сегодня был дождик. Утром идет, как будто, снежок, и скоро все тает. На реке появилась опять вода поверх льда. Погода сырая, скверная, гадкая, я чувствую себя скверно вообще в сырую погоду. Сегодня меня морозит... что-то вроде лихорадки. Пишу вот, а руки уже трясутся... Днем идет сляка, на улице грязь, сырость. Днем занимаюсь в милиции. Из Бийска приехала Таська, жена Петькина. Она была в Бийске.” (Константин Фёдорович Измайлов, 26 марта 1926 г.)*

В (1) рассказывается о реальных событиях, наблюдаемых автором - погоде, физических ощущениях, событиях в жизни людей, связанных с автором.

В (2) приводится пример записи, посвященной внутренней душевной жизни автора:

2) “Я не могу хорошенько понять, что со мной делается, в чем корни моего состояния — в психике или в физике. Не то устало сердце, не то опустела душа. Совершенно все как в тумане. Но отчаяния во мне нет. Я ясно и просто приняла факт неудачи в своей личной жизни. [...] Ничего не поделаешь. Он говорит, что у меня старомодные понятия о любви, о выражении ее. Он не прав, потому что дело не в способах выражения ее, а в сущности. <...> Прощай, Алеша! Спасибо тебе за все, что было.” (Зинаида Антоновна Денисьевская, 10.03.1930 г.)

Запись (3) иллюстрирует более редкий тип — когда записью является не фиксация внешней или внутренней жизни автора, а собственное сочинение автора или цитата из другого текста: Пример литературного текста:

3) “Написала стих о Володе Оксиковском, в начале идут эпиграфы из Володиных стихов:

Не видно, где мой причал...

Я в этой пустыне когда-нибудь сгину...

>Ветер колотит мне в грудь,

>Звёздное небо, как вина глоток

>В. Оксиковский

>Ах, какой гениальный, дитя моё,

>Матом ругающийся в минуты невзгод,

[...]

>Но не как все — гениальный.

>Можно слушать тебя часами.

>Не надоедает голос твой,

>Твоя энергия не иссякает.

Гуляла в парке, очень красива вода. Три чайки. Утки держатся стаями."
(Галина Г. Ларская, 8.10.2014 г.)

Кроме того, записи, посвященные описанию внешних событий можно разделить на два хорошо представленных типа: описание событий повседневной жизни человека (как в (1)) и описание событий, связанных с определенной деятельностью, в которой он принимает участие и за которую он ответственен (яркий пример этой категории — рабочие отчёты, как в (4)).

4) *"Сегодня из Москвы прилетели генералы А.Н. Ефимов, М.Н. Мишук, Л.И. Горегляд и Н.А. Бабийчук. Вместе с ними прилетел и летчик-космонавт Волынов с отрядом слушателей-космонавтов, все они —*

Джанибеков, Илларионов, Березовой, Романенко, Исаулов, Дедков, Козлов, Попов и Фефелов — впервые на космодроме, и им здесь есть что посмотреть. [...] Провел совещание, на котором уточнили вопросы, связанные с участием личного состава ВВС в заседаниях Госкомиссии, подготовке пусков, управлении полетом, работе службы поиска и спасения. Весь личный состав разбили на три смены для круглосуточного дежурства на КП. Руководителями смен назначили генералов Горегляда, Кузнецова и Николаева." (Николай Петрович Каманин, 18.04.1971 г.)

б) Плотность событий в тексте

В некоторых текстах последовательно перечисляются многочисленные события, случившиеся с автором записи, в других, напротив, подробно описывается только один эпизод. Под **событийной плотностью** мы подразумеваем соотношение количества этих событий и длины текста (пока этот признак определяется читательской интуицией и не формализовано строго), см. примеры (5) и (6).

5) "Мы с Яшенькой на выставке П. Кузнецова, Истомина, Волкова, Бебутовой, Щеглова и др. в Музее восточных культур. Кроме того, японские, китайские скульптуры, картины и др. Встретили Юру Злотникова и беседовали, а потом вместе пошли к Ирке в библиотеку и там смотрели мультфильмы. Далее я, Яшка, Ирка поехали к Ильке на Катенькино полугодие, и Яшка заснул по дороге у меня на плече. У Ильки он проснулся, играл с Андрейкой, Володиным сыном. Я выпил пару рюмок чачи, ел капустный пирог и пр. Было скучно. Были еще: Илька, Ирина Ник., Зуска и Володя. Катенька — милое беленькое создание без волос."

(Михаил Яковлевич Гробман)

В (5) событийная плотность высокая: в сравнительно короткой записи перечислено несколько событий.

В (6), напротив, представлен только один эпизод.

6) “Сегодня, когда мы рыли окопы в лесу, появился рассерженный Бобков. Полковой комиссар, оказывается, привез из Кронштадта приказ Военного совета о моем назначении редактором многотиражной газеты. Не застав меня на месте, он решил, что в погоне за романтикой десантной жизни я бросил газету на произвол судьбы.

— Товарищ старший политрук, немедля возвращайтесь на базу и займитесь своим прямым делом, — увидев меня, приказал Бобков. — И больше прошу без моего разрешения не покидать базу.

Не понимая, почему полковой комиссар говорит со мной таким официальным тоном, я все же решил обратиться к нему с просьбой:

— Разрешите захватить с собой и наборщика? Я без него не справлюсь.

[...]”

(Пётр Иосифович Капица, 18.07.1941 г.)

в) Функциональный стиль

Ещё одним признаком, по которым можно различать записи, является их функциональный стиль. Кроме наиболее распространенного разговорного стиля, в дневниках также можно встретить официальный, например, в отчётах о мероприятиях, см. пример (7).

7) “Состоялся Президиум Совета Министров СССР.

Слушали доклад В.Э. Дымшица по проекту распоряжения правительства об экономии топлива и электроэнергии.

В сообщении председателя комитета народного контроля А.М. Школьникова (которое касалось результатов проверки расходования топлива) было отмечено, что Минэнерго в этом году уже перерасходовало более 1 млн. т топлива на производство электрической и тепловой энергии. Заседание

президиума вел Н.А. Тихонов, который потребовал от министров (прежде всего от Минэнерго) выполнения норматива расходования топлива.” (Пётр Степанович Непорожний, 26.11.1979 г.)

Для сравнения представим пример разговорного стиля (8).

8) “Вербное Воскресение. Утром написал иокохамскому консулу М. М. Геденштрому, чтобы выхлопотал русским девочкам, воспитывающимся в католическом монастыре, позволение приехать сюда поговеть и встретить Пасху.

До Литургии крещено взрослых и детей более двадцати человек. За Литургией до ста причастников. Много христиан в Церкви, были и русские. С шести часов вечерня и повечерие по новому переводу, еще рукописному.” (Николай Японский, 20.04.2020 г.)

Приведенные выше признаки могут по-разному сочетаться между собой: например, запись с низкой событийной плотностью может описывать как эмоциональное состояние человека, так и реальные наблюдаемые события. Проанализировав наиболее типичные сочетания приведенных выше признаков в тексте, мы создали первый вариант классификации дневниковых записей, состоящий из пяти категорий(в скобках — теги, используемые для обозначения категорий) :

а) Типичный дневниковый нарратив (NAR), в котором описана последовательность событий

б) Описание деятельности (WORK), как правило, рабочей

в) Описание эмоциональных переживаний (EMO)

г) Описание отдельных эпизодов из жизни (EPISODE)

д) Литературные произведения или их фрагменты (LIT)

Рассмотрим то, как проявляются признаки в различных классах, в таблице 1.

Таблица 1. Признаки классов дневниковых записей

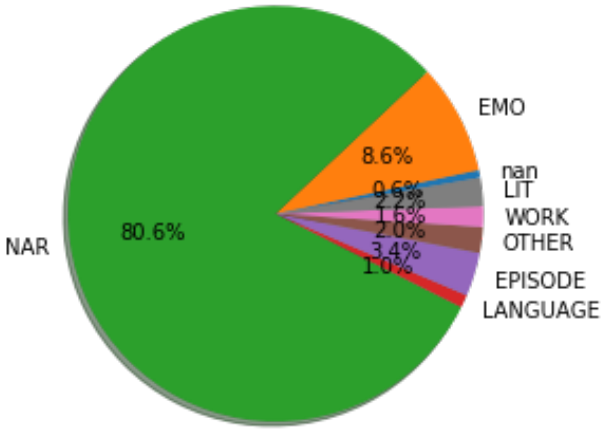
	Предмет описания	Плотность событий	Стиль
NAR	Физически происходившие события: повседневность	Высокая	Разговорный
WORK	Физически происходившие события: определенная деятельность	Средняя или высокая	Официальный, реже разговорный
EMO	Чувства, эмоции и размышления автора	Средняя или низкая	Разговорный
EPISODE	Физически происходившие события: повседневность	Низкая	Разговорный
LIT	Вымышленные события (как физические, так и нет)	Низкая	Художественный

В следующем разделе мы рассмотрим то, как выделенные выше категории представлены в корпусе дневников.

2.2.2. Анализ классификации на случайной выборке из корпуса дневниковых записей

На следующем этапе работы мы выбрали 500 случайных записей и разметили их. На рисунке 1 можно увидеть соотношение категорий.

Рисунок 1. Соотношение категорий в случайной выборке.



Кроме упомянутых выше категорий, среди тегов также есть OTHER (записи, которые сложно классифицировать) и LANGUAGE (записи на иностранных языках).

В процессе разметки мы столкнулись со следующими сложностями классификации:

а) Слишком короткие записи сложно относить к какой-либо конкретной категории (исходя из наших критериев, они могут попадать как в категорию NAR, так и в категорию EPISODE, поскольку их плотность нельзя определить однозначно):

9) “Был у герцога бал”. (Дмитрий Михайлович Волконский, 23.01.1814 г.)

б) Слишком длинные записи сложно классифицировать прежде всего из-за неравномерной событийной плотности: они часто представляют последовательность подробно описанных эпизодов (см. пример (10)). Также они могут сочетать два или более типа записей: например, отрывок из стихотворения (LIT), эмоциональное рассуждение (EMO) и короткий рассказ о прошедшем дне (NAR). В таком случае смешанный тип будет не только у дневника, а и у

отдельной записи.

10) “*Париж*.

Ночью умер Анджей Вайда. Что я помню: Мачека, хватающегося окровавленными руками за простыни, трупы рабочих в разрушенном костеле, конвейер катынского расстрела, тонущего юношу в «Аире», истерику на похоронах Цыбульского. Но еще и непостижимую педофилию «Врат рая».

Трачу сорок сонных минут, чтобы заказать билет в Гран-пале на мексиканскую выставку, сайт постоянно слетает. Зря страдаю, потому что очереди нет. Выставка грандиозная — от семейных портретов XIX века до листов молодого шарлатана, предлагающего прохожим обрисовывать контуры плиток на станции парижского метро. Ривера, которого я недолюбливал из-за коммунизма, к старости стал гениальным художником, хотя поначалу был так себе. Торговка с белыми каллами!

[...]

Возвращаемся по темному проходу под железнодорожным мостом, тут шел Мёрфи из фильма «Love» после драки с любовником Электры. Лиля боится, что в этом вульгарном районе у нее отберут айфон, и теперь мне чудится, будто все хмыри пожирают глазами мой девайс, особенно один тип, бесцеремонно ко мне присматривавшийся в вагоне и тоже вышедший на Клиши. Чтобы от него избавиться, резко меняю курс и, оказавшись на улице, погружаюсь в приступ аллергического чихания, похожего на музыку Ласкано.” (Дмитрий Борисович Волчек, 10.10.2016 г.)

в) Категория LIT кажется избыточной, особенно при автоматическом анализе. Несмотря на то, что зачастую литературный текст является частью нарратива даже у непрофессиональных авторов (например, стихи в эмоциональной записи), его сложнее анализировать. Стихотворный текст слишком значительно отличается от прозаического, а в случае, когда литературное произведение представлено в прозе, его сложно отличить от обычной дневниковой записи.

г) Корпус содержит записи, форма которых затрудняет их автоматическую классификацию. К ним относятся записи на иностранных языках, записи в дореволюционной орфографии и записи с орфографическими ошибками (см. пример (11)). Первые две категории мы решили не рассматривать в ходе исследования, а третью — оставить, поскольку они в значительно меньшей степени отличаются от грамотного русского текста.

11) “Втор[ник]. Нечего асобеного не в первый раз купался – не случилось только мы ходили с тетей Варей к вечерни и я заметил что в церкови было очень опрятно и чисто потому что ожидали Архирея но он проехал мимо”. (Николай Сергеевич Андреев, 26.05.1857)

В ходе предварительной классификации мы выявили сложности, с которыми можно столкнуться при разметке корпуса. Это повлекло за собой обновление классификации прежде всего в целях стандартизации подхода к разметке, улучшения согласия аннотаторов и результатов автоматической классификации. Изменения будут описаны в следующем разделе.

2.2.3. Предобработка корпуса. Обновление классификации.

Итак, по результатам предварительного анализа было принято решение не рассматривать некоторый типы записей. Перед разметкой данных мы решили удалить из корпуса те записи, при классификации которых могли возникнуть значительные проблемы.К ним относятся:

- записи длиннее 1500 символов и короче 500 символов

- записи на иностранных языках и в дореволюционной орфографии

- записи, содержащие длинные стихотворные фрагменты (больше пяти строчек)

После предобработки корпуса количество записей в нем сократилось с 380 тысяч до 120 тысяч: значительная часть записей была короче или длиннее

пороговых значений. Это указывает на необходимость в будущем расширить возможности классификации (например, при анализе корпуса с большим количеством длинных или коротких записей отказаться от различия по признаку событийной плотности).

Новая классификация стала включать четыре категории, а не пять: типичный дневниковый нарратив NAR, описание деятельности WORK, описание эмоциональных переживаний EMO и описание отдельных эпизодов EPISODE. В дальнейшем для удобства мы будем обозначать категории их тегами.

Важно отметить, что все категории, кроме NAR и EPISODE, имеют свой отдельный предмет описания. Категории NAR и EPISODE описывают один и тот же предмет и отличаются в первую очередь плотностью событий. Поскольку не для всех записей плотность будет однородна (например, в записи может быть описано несколько событий кратко, а один эпизод — подробно), это может вызвать сложности как для автоматической классификации, так и для разметчиков. Также в отличие от всех остальных типов записей, между этими двумя записями нельзя провести различие по ключевому признаку — предмету описания. В примере (12) мы приводим запись, сочетающую высокую и низкую событийную плотность (подробно описан эпизод с лотереей и кратко — эпизоды с зубами).

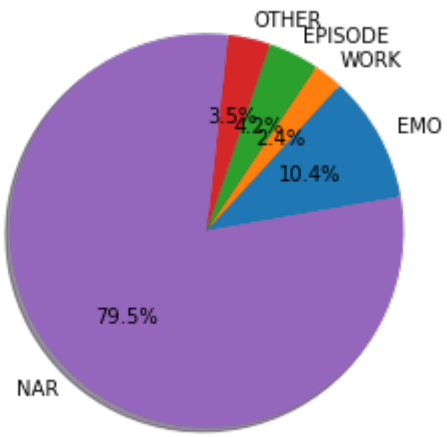
> *12) “Декабрь. 17-е число, воскресенье. Сторож женского училища приносит мне на 15-ть билетов, по 20 коп. каждый, выигрыш с лотерей в нашем училище — №8-й вещей: дамский пресс, деревянный, оклеенный бумагой, с собачкой наверху, бегущей в правую сторону по зелени, и песком внутри, вроде церковного амвона, и поздравляет с выигрышем, говоря: «Не дорога вещь, а дорого счастье». Полозову — маленькие часы, представляющие Сухареву башню в Москве, Поверенному — картины с., что подвальному, губернатору — картина Сусанина, Ивану — башмаки бархатные. \nA на 22-е число выпал последний зуб в правой стороне, качавшийся, на низу и составил 14-е число оных, в ихном хранилище у меня. Затем остался у меня один зуб коренной в левой стороне на пятницу*

22-го числа”. (Иван Васильевич Июдин, 29.12.1861 г.)

Тем не менее, мы решили на данном этапе сохранить категорию EPISODE, поскольку наиболее типичные примеры этой категории достаточно сильно отличаются от наиболее типичных примеров категории NAR.

На следующем этапе нашей работы было необходимо подготовить данные к разметке аннотаторами для исследования согласованности оценок и создания обучающего датасета. Несмотря на то, что из общего корпуса дневниковых записей уже были удалены сложные для классификации записи, случайная выборка из корпуса не смогла бы стать качественным материалом для эталона из-за значительного дисбаланса классов. Как было отмечено выше (рис. 1), к категории NAR относится подавляющее большинство записей. Этот дисбаланс сохраняется даже с учётом удаления ряда коротких записей (рис. 2).

Рисунок 2. Соотношение категорий после удаления из выборки записей и обновления классификации.



Таким образом, формирование обучающего датасета включало в себя дополнительные этапы.

На первом этапе мы создали небольшой (около шестиста записей)

сбалансированный датасет. На нём мы обучили модель логистической регрессии, в которой документы были представлены в виде векторов с помощью инструмента TfidfVectorizer[3], который учитывает важность определенных слов в контексте документа. F-мера полученной нами модели составила 0.67.

Использовав данную модель, мы предсказали классы для всего предобработанного корпуса дневниковых записей. Из каждого класса мы выбрали 2500 случайных записей (для класса WORK, в котором значительная часть записей принадлежала одному автору, мы вручную ограничили количество записей данного автора). В результате мы получили датасет из 10 тысяч записей, сбалансированность которого по нашей оценке должна быть выше, чем сбалансированность случайной выборки из корпуса.

2.2.4. Разметка корпуса

Для волонтёров, размечавших корпус, была создана инструкция с описаниями классов и примерами, а также отдельным описанием некоторых спорных случаев (приложение 3). Разметчики должны были прочесть запись, оценить, верно или неверно проставлен тег записи, и в случае ошибки указать верный тег. При этом запись анализировалась вне общего контекста дневника: у разметчиков не было доступа к дополнительной информации об его авторе. В случае, если однозначно определить категорию не удавалось, разметчики проставляли тег OTHER.

5955 записей было размечено дважды. Из них 315 записей получили тег OTHER как минимум один раз, 33 записи получили тег OTHER от обоих разметчиков (см. пример одной из таких записей (13)).

13) “### Диплом международного конкурса музыкантов завоевал дирижер из Ижевска \nДиплом лучшего дирижера 19-го Международного конкурса музыкантов привез из итальянского города

3 Здесь и далее для линейных классификаторов и TF-IDF представлений использовалась библиотека scikit-learn (https://scikit-learn.org/)

Милана ижевский маэстро Николай Роготнев. \nЭтот молодой человек, главный дирижер Театра оперы и балета, впервые порадовал общественность Удмуртии еще в прошлом году. Он удачно выступил на конкурсе в Копенгагене и получил приз Датского национального оркестра. А сейчас соперниками у Николая оказались сразу шестнадцать претендентов из самых различных стран. Постепенно круг сужался до двенадцати, а потом и до девяти. \nМузыкант, приехавший в Италию с родины Петра Чайковского, показал высокий класс. \nНа торжественном приеме у главы Правительства Удмуртии Павла Вершинина маэстро поделился планами на будущее. Через год он мечтает участвовать в международном конкурсе с Национальным симфоническим оркестром Болгарии". (Альфред Александрович Артамонов, 17.09.1995 г.)

Дневник журналиста Альфреда Артамонова[4] является корпусом его публикаций и выделяется из общего дневникового корпуса. Записи, в которых автор пересказывает новости и иные внешние события, вообще вызывают трудности (см. (14)).

14) *"В Венгрии продолжается наступление наших войск. Финляндия и Румыния что-то не очень хорошо выполняют условия перемирия. Во всяком случае, их приходится тащить через пень колоду. Болгария поставила у власти левые элементы и там, как будто, все в порядке. Тени неприятных ситуаций мелькают на страницах газет все чаще. [...]". (отрывок из записи Сергея Гавриловича Юрова, 8.12.1944 г.)*

Возможно, в дальнейшем стоит включить в классификацию новую категорию, которая будет объединять подобные записи. Даже если исключить дневники журналистов, записи о новостях всё равно составляют ощутимую часть корпуса — например, если речь идёт о фронтовых сводках.

Рассмотрим согласованность записей, которые оба разметчика отнесли к

4 https://prozhito.org/person/505

какой-либо из четырёх основных классификаций, на рисунках 3 и 4.

Рисунок 3. Согласованность оценок (абсолютные числа).

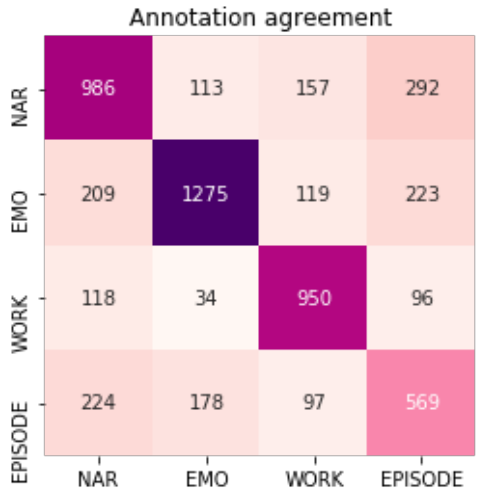
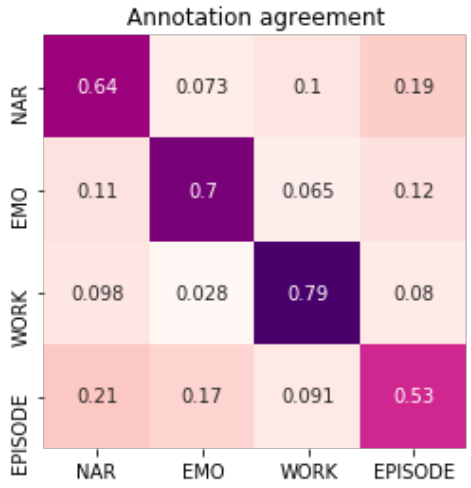


Рисунок 4. Согласованность оценок (относительные числа).



Можно заметить, что наибольшая согласованность оценок относится к категории WORK, а наименьшая — к категориям NAR и EPISODE.

Рассмотрим количество записей, получивших разные оценки:

- 516 — NAR и EPISODE

- 401 — EMO и EPISODE

- 322 — NAR и EMO

- 275 — NAR и WORK

- 193 — WORK и EPISODE

- 153 — WORK и EMO

Исходя из этого, можно сделать вывод, что наименее четкие границы между категориями свойственны категориям NAR и EPISODE, а также EMO и EPISODE. Наиболее чётко разделены категории WORK и EMO, но и для них существуют спорные случаи.

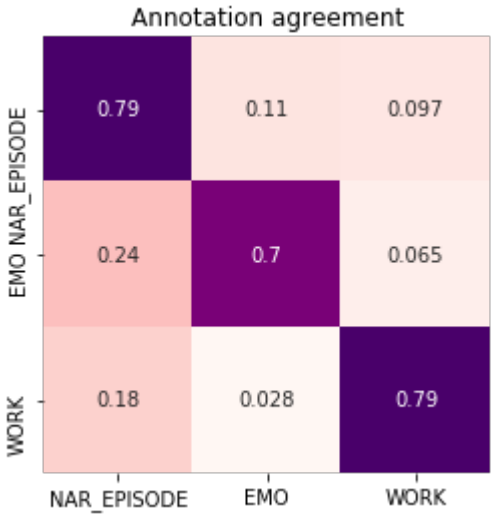
Например, запись из дневника общественного деятеля XIX века Александра Половцова, в которой тематика работы связана с эмоциональной оценкой:

> *15) “Пятница. Гр. Протасов-Бахметев назначен главноуправляющим Собственною канцеляриею по делам императрицы Марии. Это прекрасный в нравственном смысле человек, безукоризненно честный, правдивый, добросовестный, но при весьма ограниченном кругозоре, не имеющий понятия ни о государственном, ни о каком бы то ни было ином управлении. Он весь пропадет на мелочах, а большими делами будет вертеть первый, кто им овладеет и притом весьма легко овладеет. \ nВечером в клубе разговор с Абазою об обуховском деле, подлежащем докладу на следующий день. Абаза находит единственно правильным основанием к разрешению этого дела удовлетворение ходатайства Николая Левашева о возвращении пайщиками истраченной ими на покупку паев суммы.” (Александр Александрович Половцов, 30.05.1890)*

Попробуем объединить категории NAR и EPISODE в одну (рисунок 5). В этом случае согласованность значительно возрастает. Это подтверждает предположение о том, что категории NAR и EPISODE достаточно схожи, а

провести чёткую границу между ними сложно.

Рисунок 5. Согласованность оценок (относительные числа) для 3 классов



Итак, выше был описан процесс подготовки классификации для автоматического анализа дневниковых записей и апробация этой классификации на разметке обучающего датасета. Мы создали собственную классификацию дневниковых записей. На наш взгляд, полученная классификация в целом удовлетворяет требованиям, которые мы к ней предъявляем. Рассмотрим подробнее, что удалось сделать, а с чем возникли трудности.

Требование 1. Основная единица классификации — дневниковая запись.

Требование соблюдено. Единственная сложность здесь заключается в том, что некоторые записи (особенно длинные) сложно классифицировать однозначно из-за их неоднородной структуры. На данном этапе анализа этим можно пренебречь, а в дальнейшем можно, например, выделять для такой записи две категории: основную и дополнительную

Требование 2. Классификация должна быть универсальна.

Требование соблюдено. Из 5955 записей, случайным образом выбранных из общего корпуса, только 315 хотя бы один раз были отнесены в категорию “другое”. Это показывает, что подавляющее большинство записей можно отнести к одной из категорий классификации. Тем не менее, классификацию можно было бы дополнить, добавив категорию для пересказа новостей и прочих событий, в которых рассказчик не принимал участия.

Требование 3. Критерии, по которым определяется категория, прозрачны.

Здесь в процессе разметки были выявлены недостатки критериев. Согласованность оценок аннотаторов достаточно высока, но всё ещё существует ряд спорных случаев, которые с трудом поддаются классификации. Каждый из трёх значимых признаков (событийная плотность, предмет описываемых событий и стиль текста) может быть двояко интерпретирован для подобных случаев. Тем не менее, мы предполагаем, что добиться идеальной прозрачности критериев невозможно из-за особенностей исследуемых нами данных.

Выше мы неоднократно приводили примеры смешения категорий относительно критерия плотности событий при сравнении категорий NAR и EPISODE. Для признака “реальность” спорным случаем может быть, например, размышление автором записи о поступке его приятеля (к описанию реального события добавляется эмоциональная оценка), а для признака “стиль” — не слишком строгий отчёт о работе с использованием разговорных слов.

Таким образом, мы считаем, что разработанная нами классификация дневниковых записей удовлетворяет нашим требованиям и достаточна для разработки на её основе моделей автоматической классификации записей. Созданию этих моделей будет посвящена следующая часть исследования.

3. Создание моделей автоматической классификации дневниковых записей

Несмотря на то, что для каждой из представленных выше категорий можно выделить определенные лексические особенности, мы предполагаем, что между ними существуют и другие различия, в том числе стилистические. Поэтому при представлении текстов дневниковых записей в векторном виде мы будем использовать как классические подходы, основанные на лексическом составе текста (TF-IDF векторизация, эмбеддинги), так и иные признаки (частеречные теги, частотность тех или иных знаков пунктуации).

В большинстве экспериментов мы будем использовать классификацию, которая содержит все четыре категории классификации (NAR, EPISODE, EMO, WORK). Для наиболее успешных моделей мы проверим их качество на классификации с тремя категориями, в которой NAR и EPISODE будут совмещены. Мы предполагаем, что, поскольку интенция этих категорий очень схожа (основной признак, предмет описания, совпадает), объединение позволит нам повысить качество классификации и при этом не потерять важную информацию о нарративном поведении автора.

3.1. Подготовка датасета

На основе тех записей, для которых мнение разметчиков совпало, мы сформируем два обучающих датасета: в один из них будут входить все записи нашего “золотого стандарта”, из другого мы удалим часть записей, чтобы сбалансировать датасет . Несбалансированный датасет состоит из 3780 записей, соотношение категорий EMO/NAR/WORK/EPISODE = 1275/986/950/569. В сбалансированный входит 2240 записей (560 записей для каждого класса).

3.2. Использование лексических признаков

Прежде всего, мы обратились к экспериментам, опирающимся на на лексические особенности текстов разных категорий. Сначала мы использовали

линейные классификаторы (логистическая регрессия, наивный байесовский классификатор, метод опорных векторов), а затем — нейронные сети.

3.2.1. Анализ ключевых слов текста

Перед разработкой моделей мы извлекли ключевые слова, характерные для разных категорий текстов, с помощью метода TF-IDF. Мы не лемматизировали слова в текстах, поскольку представлялось интересным проанализировать также различные формы, в которых слово может встречаться в документах. При анализе мы удалили из корпуса стоп-слова для русского языка (источник стоп-слов — библиотека Natural Language Toolkit [ссылка на nltk]).

Таблица 2. Ключевые униграммы для каждой из категорий.

	EMO	EPISODE	NAR	WORK
1	то	то	сегодня	ссср
2	очень	очень	очень	квт
3	сегодня	сегодня	com	минэнерго
4	время	время	id	работы
5	жизни	день	день	цк
6	com	вчера	то	энергетики
7	id	говорит	вечером	сегодня
8	жизнь	нам	утром	15
9	вчера	сказал	время	очень

10	день	около	часов	10
11	всё	несколько	дома	оборудования
12	что	утра	домой	аэс
13	человек	вечером	вчера	вопрос
14	кажется	лет	10	гэс
15	могу	который	нам	млн
16	хотя	человек	го	совета
17	по	ночь	днем	строительства
18	людей	наши	дня	время
19	нибудь	утром	утра	бр
20	знаю	часов	погода	12

Таблица 3. Ключевые би- и триграммы для каждой из категорий.

	EMO	EPISODE	NAR	WORK
1	com id	com id	com id	com id
2	что то	какой то	весь день	цк кпсс
3	как то	что то	из за	млн квт
4	все таки	кто то	что то	тыс квт

5	из за	из за	утром встал	министров ссср
6	какой то	друг друга	какой то	совета министров
7	почему то	как то	как то	совета министров ссср
8	друг друга	где то	каждый день	бр игады
9	сих пор	все таки	пили чай	млн руб
10	где то	кое где	днем занят	развития энергетики

По приведенным выше таблицам можно заметить, что категории EPISODE и NAR действительно близки (11 общих ключевых слов из 20 униграмм, 5 общих из 10 би- и триграмм). Также значительно связаны категории EPISODE и EMO (7 и 8 (!) пересечений соответственно), EMO и NAR (8 и 5 пересечений).

Категория WORK кажется наиболее независимой от других лексически. Тем не менее, при выделении ключевых слов в ней есть свои сложности. Значительную часть корпуса (даже учитывая ограничение их количества при формировании выборки) составляют достаточно однотипные записи министра энергетики Петра Степановича Непорожнего, которые во многом и сформировали список ключевых слов.

Другая важная особенность получившихся списков — то, как на эти списки повлиял формат анализируемых данных. В корпусе дневников “Прожито”, во-первых, присутствуют ссылки на некоторых из участников описываемых событий, и, во-вторых, часто используются сокращения. Это объясняет присутствие в списках соответственно токенов “com” и “id”, являвшихся частью ссылок, и таких единиц как “го” (по всей видимости, сокращение от порядковых числительных, например, “14-го”) и “бр” (распространенное в одном из “рабочих” дневников сокращение для “бр[игада]”).

Мы не стали проводить дополнительную предобработку корпуса, удаляя все ссылки и исправляя сокращения. Мы считаем, что исследование того, как эти единицы используются в тексте, тоже может быть важным для понимания структуры записей (например, можно заметить, что в категориях EMO и NAR упоминания людей (токен “id”) более значимы, чем в категориях EPISODE и WORK). Возможно, более пристальное внимание к особенностям формата корпуса “Прожито” и более тщательная его предобработка может быть полезна в будущих исследованиях данного корпуса.

3.2.1. Использование линейных классификаторов для классификации дневниковых записей

На первом этапе мы использовали TF-IDF представления и линейные классификаторы: такие, как логистическая регрессия (LogReg), наивный байесовский классификатор (NaiveBayes) и метод опорных векторов (SVM). При создании моделей использовалась библиотека scikit-learn.

Ниже в таблицах 4 и 5 можно увидеть результаты работы моделей на сбалансированном и несбалансированном датасете. В таблице представлены результаты для упомянутых выше классификаторов с разными размерами n-грамм. Качество несколько увеличивается при добавлении к униграммам биграмм, но добавление следующих n-грамм практически никак не влияет на качество. Также качество значительно падает, если исключить униграммы.

При анализе мы также рассмотрели и другие гиперпараметры. Описанные ниже модели (кроме модели биграмм и триграмм) обладают следующими параметрами:

- min_df (минимальное количество документов, в которых есть n-грамма) = 10

- max_df (максимальное количество документов, в которых есть n-грамма) =

90%

При незначительном изменении этих параметров качество работы классификатора практически не менялось, при большем отклонении от них — падало.

Таблица 4. Качество моделей на сбалансированном датасете для 4 категорий.

модель	f-measure
TF-IDF (1,1) + LogReg	0.80
TF-IDF (1,2) + LogReg	**0.81**
TF-IDF (1,5) + LogReg	**0.81**
TF-IDF (2,3, min_df=1) + LogReg	0.60
TF-IDF (1,1) + NaiveBayes	0.76
TF-IDF (1,2) + NaiveBayes	0.76
TF-IDF (1,5) + NaiveBayes	0.76
TF-IDF (1,2) + SVM	0.77

Таблица 5. Качество моделей на несбалансированном датасете с 4 категориями.

модель	f-measure (micro)	f-measure (macro)

TF-IDF (1,1) + LogReg	0.79	**0.76**
TF-IDF (1,2) + LogReg	**0.80**	**0.76**
TF-IDF (1,5) + LogReg	**0.80**	**0.76**
TF-IDF (2,3) + LogReg	0.47	0.37
TF-IDF (1,1) + NaiveBayes	0.72	0.61
TF-IDF (1,2) + NaiveBayes	0.72	0.62
TF-IDF (1,5) + NaiveBayes	0.72	0.62
TF-IDF (1,2) + SVM	0.77	0.69

Можно заметить, что логистическая регрессия показывает чуть лучшие результаты, чем остальные модели, а также менее чувствительна к дисбалансу данных.Таким образом, за базовое решение было решено принять модель TF-IDF (1,2) + LogReg.

Рассмотрим более подробные результаты её работы на обоих датасетах (рисунки 6 и 7).

Рисунок 6. Матрица ошибок базового решения на сбалансированном датасете.

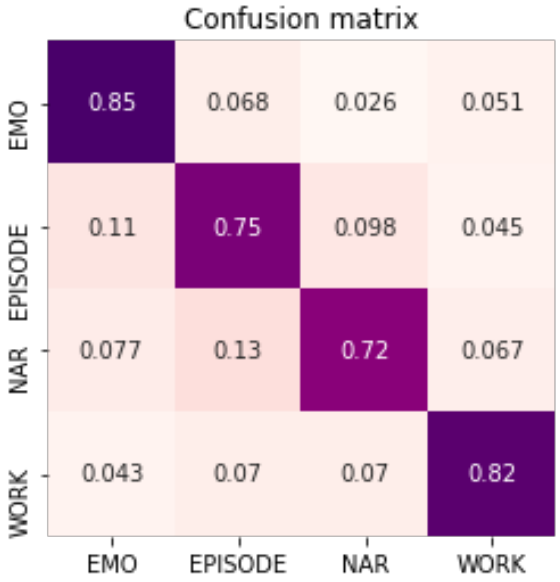
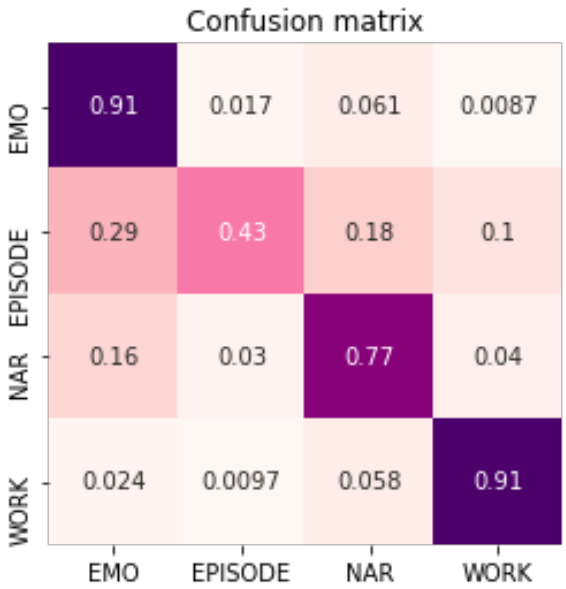


Рисунок 7. Матрица ошибок базового решения на несбалансированном датасете.



Основная разница при анализе двух датасетов видна при анализе категории EPISODE. F-мера для всех категорий отличается незначительно, но сбалансированный датасет показывает меньшую разницу в качестве между

категориями, чем несбалансированный.

При этом для обоих датасетов можно проследить общие закономерности: лучше всего предсказываются категории WORK и EMO. Это, по-видимому, связано с тем, что в них проще всего выделить специфические темы и характерную лексику (профессионализмы и сокращения в первом случае и слова, описывающие мысли и чувства, во втором).

Проанализируем работу модели на новых датасетах: сбалансированном и несбалансированном с тремя категориями (WORK, EMO, NAR_EPISODE).

Мы использовали тот же датасет, состоящий из записей, для которых совпадало мнение разметчиков. Несбалансированный датасет состоял из 3024 записей обучающей выборки и 756 записей тестовой выборки (1555 записей для категории NAR_EPISODE, 1275 для категории EMO и 950 для категории WORK). В сбалансированном каждая категория была представлена 950 записями, объём обучающей выборки составил 2280 записей, а объём тестовой — 570.

F-мера для сбалансированного датасета составила 0.86, для несбалансированного — 0.88 (как для микро-, так и для макроусреднения). Результаты подтвердили нашу гипотезу о том, что объединение похожих категорий положительно влияет на качество классификации.

Рисунок 8. Матрица ошибок базового решения на сбалансированном датасете, 3 класса.

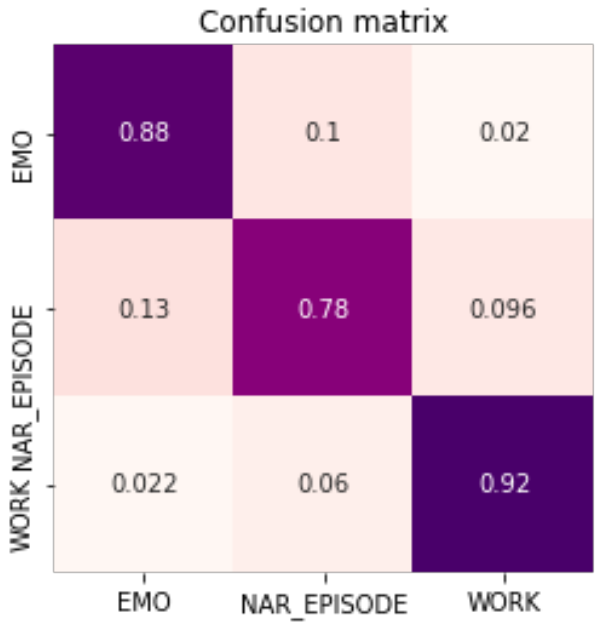
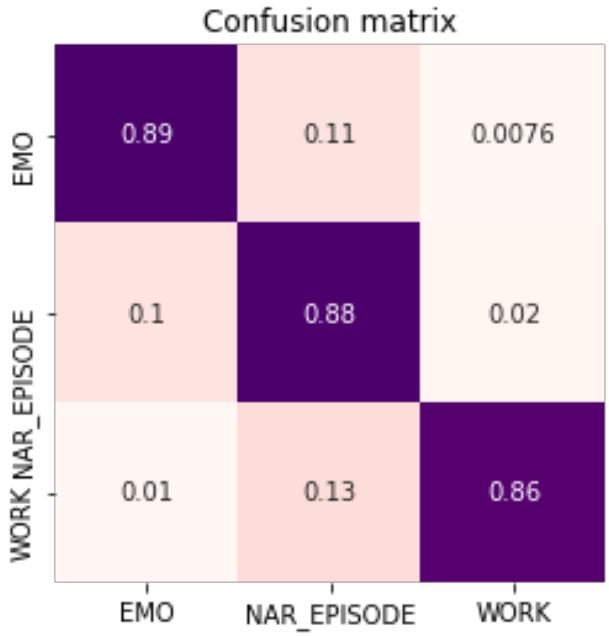


Рисунок 9. Матрица ошибок базового решения на несбалансированном датасете, 3 класса.



Мы рассмотрели применение различных линейных классификаторов к решению нашей задачи. Лучшие результаты показала модель TF-IDF (1,2) + LogReg. В следующем разделе мы рассмотрим применением к нашим данным нейронных сетей.

3.2.2. Использование нейронных сетей для классификации дневниковых записей

На следующем этапе мы использовали нейронные сети. В экспериментах мы сочетали несколько видов представления текста (частотные слова, TF-IDF, обучаемые эмбеддинги, предобученные эмбеддинги) и две архитектуры нейронных сетей (сеть прямого распространения, свёрточная нейронная сеть).

Мы использовали предобученные эмбеддинги ruwikiruscorpora_upos_skipgram_300_2_2019, загруженные с сервиса RusVectores[5] и обученные на Википедии и Национальном корпусе русского языка. Из 27,7 тысяч токенов, содержавшихся в обучающей выборке сбалансированного датасета, для 6,5 тысяч предобученные эмбеддинги отсутствовали. Например, не было векторных представлений для имен собственных (“димушка”, “в.сорокин”) и специфических терминов(“мухоловка-пеструшка”, “кернохранилище”, “высокоудойный”).

Результаты классификации можно увидеть в таблице 6.

Таблица 6. Качество классификации с использованием нейронных сетей.

	f-мера (датасет сбалансирован)	f-мера micro (датасет не сбалансирован)	f-мера macro (датасет не сбалансирован)
Freq + Feedforward	0.80	0.81	0.79
TF-IDF + Feedforward	0.77	0.77	0.75

5 https://rusvectores.org/ru/

Trainable embeddings + Feedforward	0.69	0.74	0.69
Trainable embeddings + CNN	0.77	0.78	0.76
Ruwikiruscorpopa embeddings + Feedforward	0.65	0.65	0.60
Ruwikiruscorpopa embedding + CNN	0.77	0.77	0.73

Использование нейросетей и предобученных эмбеддингов не улучшило качество классификации существенно. Мы предполагаем, что это может быть связано с небольшим объёмом датасета и особенностями лексического состава текстов (в частности, отсутствием предобученных эмбеддингов для многих лексем). Возможно, в будущем было бы полезным обучить эмбеддинги на всём корпусе “Прожито”.

3.3. Использование прочих признаков

На следующем этапе исследования мы выделим новые признаки, которые могут указывать на категорию дневниковой записи. Мы рассмотрим такие признаки, как частотность различных частей речи, знаков препинания и других символов, а также удобочитаемость дневниковых записей. В качестве классификатора будет использована логистическая регрессия.

3.3.1. Генерация признаков

Мы выделили следующие признаки, которые могут указывать на категорию

записи:

1. Удобочитаемость

К этим признакам мы относим среднюю (mean) и медианную длины предложений (в словах) для текстов, среднюю длину слова в слогах, а также метрику удобочитаемости Флеша-Кинкейда, рассчитанную по формуле для русского языка (Оборнева 2005).

2. Частотность знаков препинания и прочих символов

Частотность некоторых знаков препинания в предложениях. Также мы включили в эти признаки знак градуса (встречающийся во многих записях категории NAR при измерении температуры).

3. Частотность частеречных тегов

Частотность частей речи в тексте. Для определения частей речи мы использовали морфологический анализатор pymorphy2[6].

3.3.2. Результаты использования новых признаков

Проверив работу моделей на сбалансированном датасете, мы обнаружили, что качество классификации значительно уступает классификации с использованием лексических признаков. Поэтому мы приняли решение не проводить отдельные эксперименты для несбалансированного датасета.

Таблица 7. Качество классификации с использованием новых признаков.

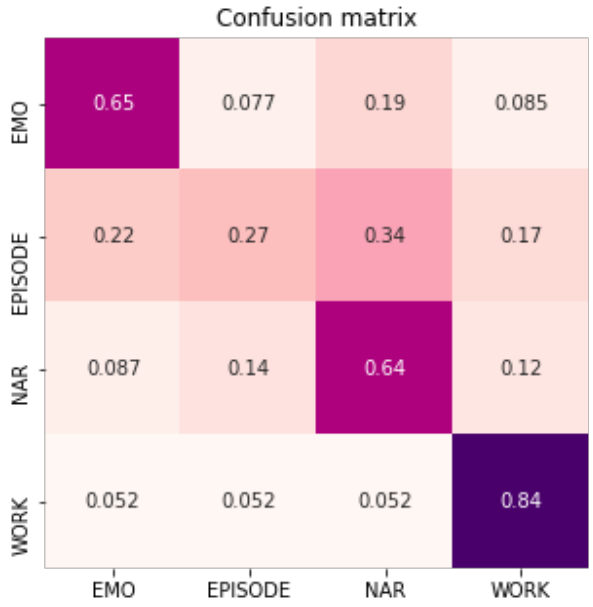
	F-мера
Readability + LogReg	0.40
(Readability + Punct) + LogReg	0.49

6 https://pymorphy2.readthedocs.io/en/latest/

(Readability + Punct + PoS) + LogReg	**0.58**

Качество классификации росло по мере добавления новых признаков. Тем не менее, модель, не использующая лексические признаки, показывает не слишком высокие результаты. Возможно, в будущем результаты можно будет улучшить, добавив новые признаки, например, выявив типичные для разных категорий дискурсивные маркеры.

Рисунок 10. Матрица ошибок модели дополнительных признаков. Сбалансированный датасет.



Рассмотрим матрицу ошибок. Можно сделать вывод, что WORK отличается от других категорий не только специфической лексикой, но и стилистическими признаками. Для EMO, например, лексические признаки кажутся более значимыми. Категория EPISODE по-прежнему выделяется хуже остальных.

Итак, выше мы описали результаты ряда экспериментов по автоматической

классификации дневниковых записей. Лучшие результаты были показаны моделью TF-IDF (1,2) + LogReg (представление TF-IDF, использующее униграммы и биграммы, и логистическая регрессия) и моделью Freq + Feedforward (представление, основанное на частоте слов, и модель прямого распространения) . Модели показали F-меру 0.81 и 0.80 соответственно при классификации записей на четыре категории на сбалансированном датасете.

F-мера модели TF-IDF (1, 2) + LogReg для трёх категорий на сбалансированном датасете составила 0.86, на несбалансированном — 0.88.

Мы предполагаем, что в будущем качество классификации может быть улучшено, во-первых, увеличением обучающего датасета, во-вторых, добавлением новых признаков (например, дискурсивных маркеров), и в-третьих, использованием более передовых архитектур нейросетей.

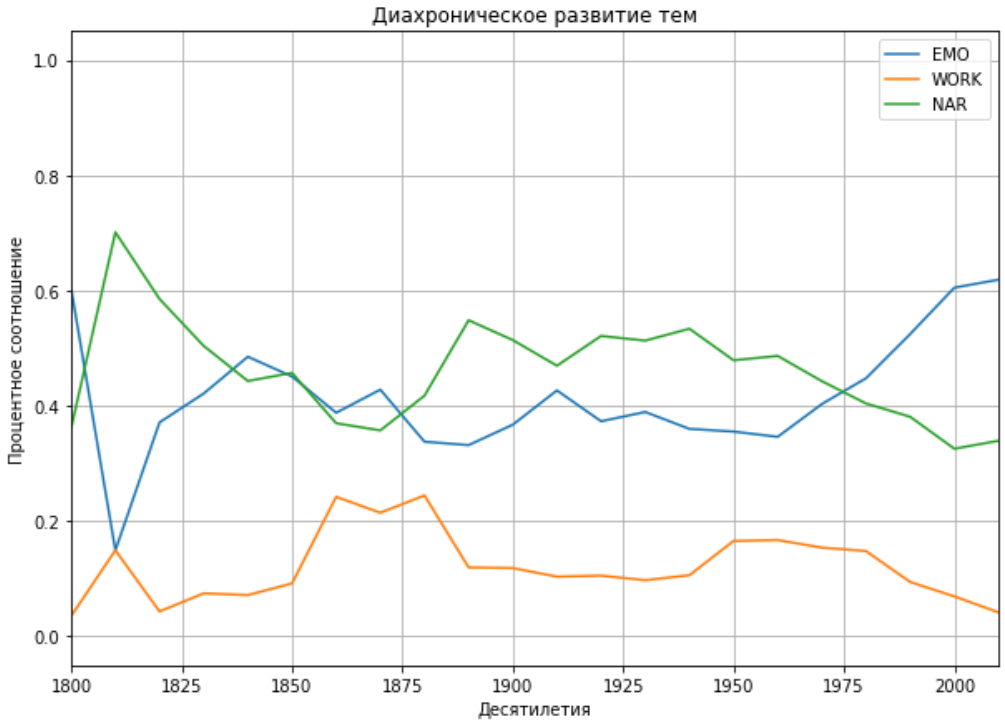
4. Использование методов дальнего чтения для анализа корпуса дневников

Завершающий этап исследования посвящен попытке с помощью разработанного алгоритма выявить особенности поведения авторов дневников на большом корпусе.

Мы обучили модель (TF-IDF (1, 2) + LogReg) на датасете, объединяющем обучающую и тестовую выборки, и предсказали тег для каждой из записей общего датасета (исключив записи на иностранных языках, в дореволюционной орфографии и записи короче 300 знаков). Для анализа мы использовали систему из трёх тегов, поскольку она легче формализуется и показала более высокие результаты при тестировании.

Первым шагом по исследованию корпуса стал диахронический анализ процентного соотношения записей в корпусе (рис. N) в период с 1800-е по 2010-е годы. Мы рассчитали процент записей для каждой из трёх категорий в каждом из десятилетий.

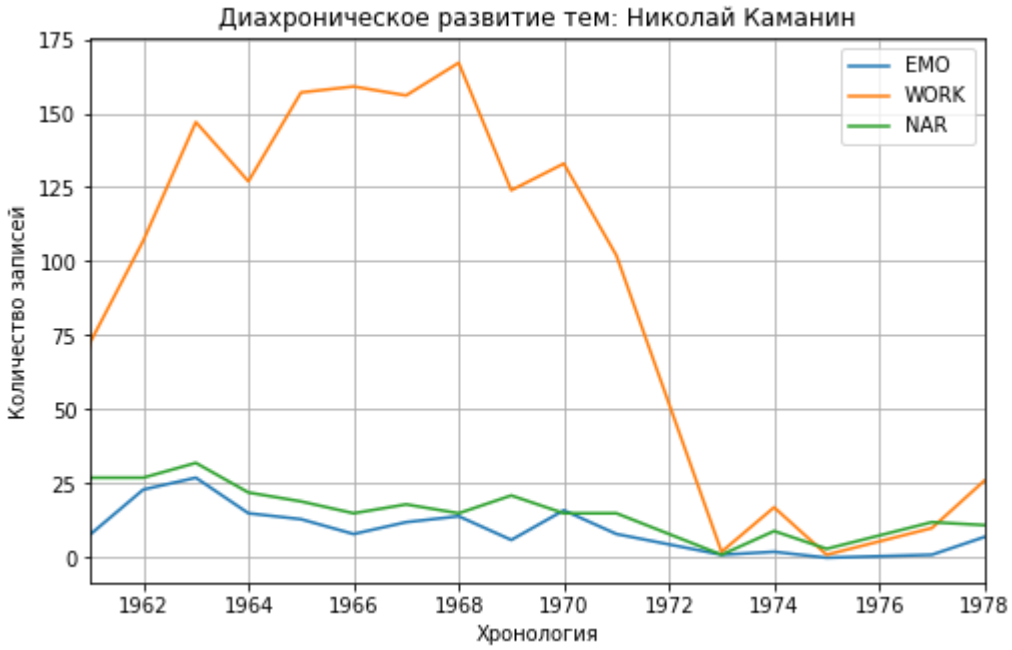
Рисунок 11. Изменение доли тем в корпусе с 1800-е по 2010-е годы.



Вероятно, изменения в соотношениях тем связаны в том числе со спецификой дневников, составляющих корпус “Прожито”. Например, во второй трети девятнадцатого века виден рост категории рабочих отчетов. Можно предположить, что он связан с большим количеством дневников государственных деятелей того времени. В связи с этим в будущем кажется интересным проанализировать, например, то, как социальное положение автора связано с его нарративным поведением.

Рассмотрим примеры, в которых мы можем связать эпизоды биографии людей с их нарративным поведением в дневниках.

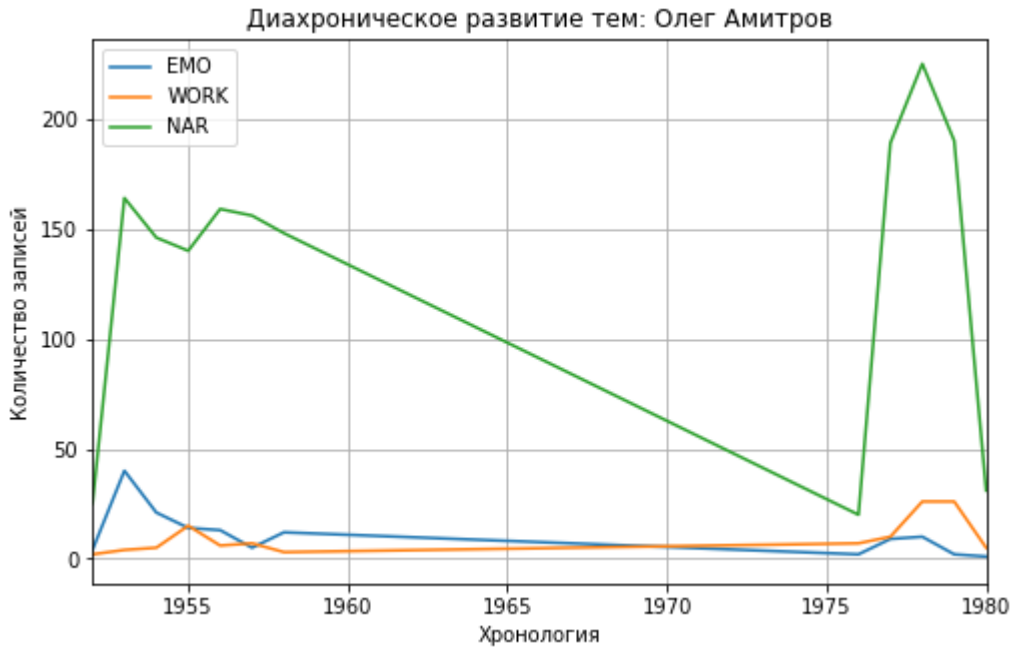
Рисунок 12. Нарративное поведение Николая Каманина.



Лётчик и военачальник Николай Каманин[78] перестаёт активно писать дневник в начале 1970-х. Особенно заметно падает количество записей, посвященных рабочей деятельности Каманина. По-видимому, это связано с отставкой лётчика в 1971 году.

7 Страница Николая Каманина в Википедии: https://ru.wikipedia.org/wiki/Каманин,_Николай_Петрович
8 Страница Николая Каманина на “Прожито”: https://prozhito.org/person/39

Рисунок 13. Нарративное поведение Олега Амитрова.



Дневник палеонтолога Олега Амитрова[9] описывает почти три десятилетия его жизни. Можно заметить, что в середине 1950-х есть небольшой пик записей категории “WORK”, что можно связать с учёбой в университете. Интересно, что при подробном рассмотрении записей категории “WORK” того времени мы заметили не только верно определенные записи (например, о докладах или семинарах), но и записи, ошибочно отнесенные к этой категории из-за тематики, связанной с искусством (см. (16)).

> *16) “Был с Игорем Ванчуровым в Большом зале консерватории на концерте молодых исполнителей. Программа была хорошая: I концерт для фортепьяно с орк[естром]., концерт для скрипки с оркестром и вариации на тему Рококо Чайковского.*

На ф-к катались на лыжах” (Олег Владимирович Амитров, 2.12.1955 г.)

9 Страница Олега Амитрова на “Прожито”: https://prozhito.org/person/3898

Данная ошибка показывает недостаток созданной нами модели. В обучающем корпусе среди записей категории “WORK” было значительное количество текстов директора Императорских театров Владимира Теляковского, вероятно, повлиявших отчасти на связь культурных мероприятий и рабочих отчётов. При дальнейшем пополнении обучающего корпуса нужно будет уделить внимание не только сбалансированности разных категорий, но и тому, как представлены в нём записи разных авторов.

Также в дневнике заметен “всплеск” эмоциональных записей в юности автора. Мы предположили, что подобное нарративное поведение в юношеском возрасте может быть характерны для многих авторов дневников. Поэтому следующим этапом нашего исследования стал анализ нарративного поведения трёх групп авторов:

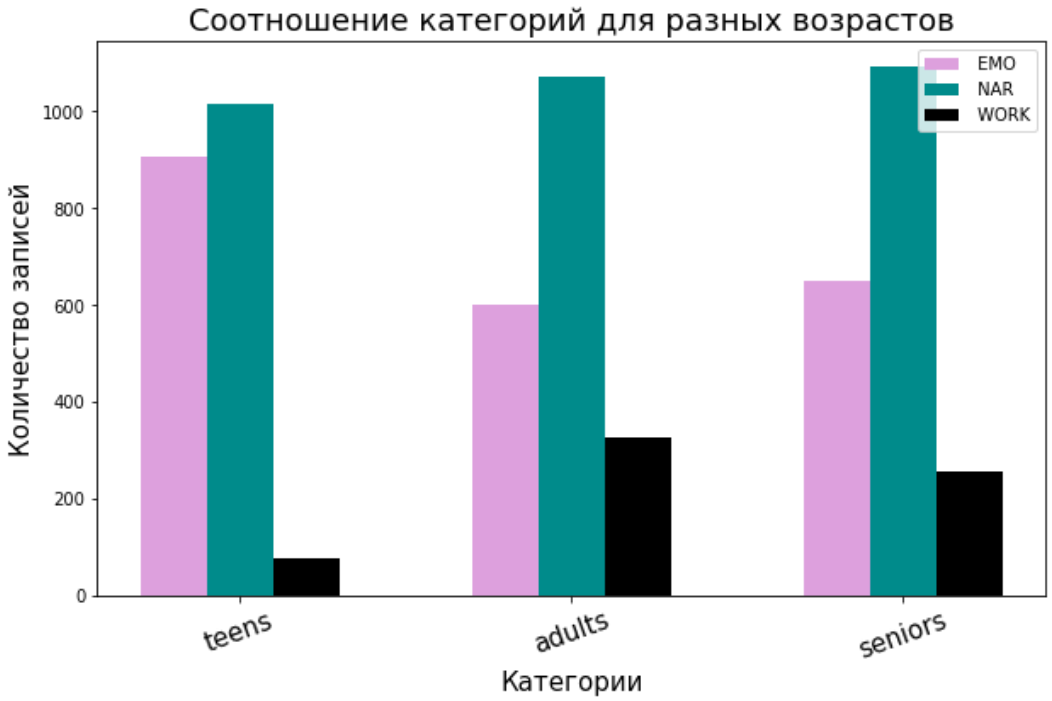
— от 13 до 20 лет

— от 30 до 40 лет

— от 70 до 95 лет

Из каждой категории мы случайным образом выбрали 2000 записей, написанных позже 1900 года. Ожидалось, что молодые авторы будут чаще писать о чувствах и реже — о работе. Также мы предполагали, что количество записей о работе снизится для категории пожилых людей.

Рисунок 14. Соотношение разных категорий в дневниках авторов разного возраста.



Можно заметить, что наше предположение относительно нарративного поведения молодёжи оказалось верным. Тем не менее, спад в категории “WORK” оказался не таким значительным, как мы предполагали: нарративное поведение пожилых авторов практически не отличается от поведения людей среднего возраста.

На наш взгляд, последнее может быть связано с тем, что значительную часть корпуса дневников составляют дневники людей, занимающихся умственным трудом, а также высокопоставленных персон. Среди авторов записей класса “WORK” в этой возрастной категории, например, балетмейстер Мариус Петипа, историк Валентин Шелохаев и министр энергетики Пётр Непорожний. Вероятно, для этих авторов достижение определенного возраста не означало прекращения рабочей деятельности.

В этой главе мы провели первые эксперименты в дальнем чтении корпусов дневников. Даже обученная на небольшой выборке размеченных данных модель

помогает выявить закономерности в содержании дневниковых корпусов. В заключении мы рассмотрим дальнейшие перспективы по развитию нашей модели и применению её для задач дальнего чтения.

Заключение

Целью работы было научиться автоматически выявлять общие закономерности в дневниках разных авторов. Для этого в нашей работе мы предложили новый формализм для описания дневника, классификацию текстов дневников на основе интенций, и провели эксперименты по автоматическому определению категории дневниковой записи.

С помощью предложенной нами формальной модели мы также проанализировали корпус дневников “Прожито”, в частности, выявили различия между особенностями нарративного поведения подростков и взрослых авторов. На наш взгляд, проделанная работа способна расширить текущие возможности дальнего чтения корпусов дневниковых записей на русском языке.

Сейчас можно отметить сразу несколько направлений, в которых можно в дальнейшем развить это исследование. Во-первых, это доработка существующей классификации, во-вторых, улучшение автоматического определения категории записи, и в-третьих, продолжение экспериментов, связанных с дальним чтением.

Доработка существующей классификации включает возможное добавление новых классов (к примеру, класса “новостей”) и объединение некоторых старых. Для разных задач можно использовать разные варианты классификации, более или менее подробные. Также в будущем можно уточнить критерии, важные для определения категории, к которой относится дневниковая запись.

Эксперименты по улучшению автоматического определения категории могут затрагивать практически каждый из этапов работы. Во-первых, представляется интересным провести эксперименты, связанные с предобработкой текста корпуса. Во-вторых, остался неохваченным ряд признаков, который может

быть важен для классификации текстов (например, дискурсивные маркеры или тональность записи). В-третьих, можно использовать более передовые модели представления текста и классификации текстовых документов (особенно в том случае, если появится больше размеченных данных).

Последнее из направлений — эксперименты с дальним чтением — является, на наш взгляд, наиболее широким. В данном исследовании представлены только первые шаги по анализу нарративного поведения в дневниках. Тем не менее, они уже показали, что данная классификация может использоваться для исследования связи нарративного поведения автора дневника и событий, происходивших в его жизни. Также возможны эксперименты по изучению закономерностей поведения целых групп людей и выделения типичного и нетипичного в дневниковых записях разных периодов. Мы ожидаем, что сочетание автоматических методов дальнего чтения с последующим “медленным чтением” дневников позволит лучше изучить особенности дневника как жанра.

Список литературы

1. Егоров 2011 — О.Г. Егоров. Русский литературный дневник XIX века. – Флинта, 2011.
2. Жожикашвили 2003 — С. В. Жожикашвили. Дневник // Литературная энциклопедия терминов и понятий (гл. ред. А.Н. Николюкин). М. 2003, с. 232.
3. Зализняк 2010 — А.А. Зализняк. Дневник: к определению жанра // Новое литературное обозрение, 6. 2010. С. 162-181.
4. Михеев 2006 — М.Ю. Михеев. Дневник в России XIX-XX века — эго-текст, или пред-текст // URL: http://uni-persona.srcc.msu.su/site/research/miheev/kniga.htm (дата обращения: 3.06. 2020). 2006.
5. Оборнева 2005 — И.В. Оборнева. Автоматизация оценки качества восприятия текста // Вестник Московского городского педагогического университета. Серия: Информатика и информатизация образования, 5. 2005. – С. 86-91.
6. Пропп 1928 — В.Я. Пропп. Морфология сказки // Гос. ин-т истории искусств. — Л.: Academia, 1928. — 152 с. — (Вопр. поэтики; Вып. XII).
7. Al Shboul 2018 — O. Al Shboul. Discourse Analysis of Refugees' Narratives: Content and Structure // Rule of Law, Courtroom Procedures, 2018. P. 51–61.
8. Aurnhammer 2019 — Christoph Aurnhammer, Iris Cuppen, Inge van de Ven, Menno van Zaanen. Manual Annotation of Unsupervised Models: Close and Distant Reading of Politics on Reddit // DHQ: Digital Humanities Quarterly 13, 3, 2019.
9. Cohen 1960 — J. Cohen. A coefficient of agreement for nominal scales // Educational and Psychological Measurement. 20 (1), 1960. P. 37–46.
10. Fischer 2016 — Fischer F., Göbel M., Kittel C., Kampkaspar D., Trilcke P. Distant-Reading Showcase. 200 Years of Literary Network Data at a Glance // Proceedings of DHd2016. 2016.
11. Labov, Waletzky 1967 — William Labov, Joshua Waletzky. Narrative Analysis: Oral Versions of Personal Experience) // Essays on the Verbal and Visual Arts. Seattle, WA: University of Washington Press, 1967. P. 12–44.
12. Li et al. 2018 — B. Li, B. Cardier, T. Wang, F. Metze. Annotating high-level structures of short stories and personal anecdotes. // Proceedings of the 11th

Language Resources and Evaluation Conference. 2018.

13. Lehnert 1981 — W. G. Lehnert. Plot units and narrative summarization // Cognitive science, 5(4) 1981. P. 293–331.

14. Moretti 2000 — F. Moretti. Conjectures on World Literature // New Left Review 1, 2000. P. 54–68.

15. Sims et al. 2019 — M. Sims, J. H. Park, D. Bamman. Literary Event Detection // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. – P. 3623-3634.

16. Rahimtoroghi et al. 2017 — E. Rahimtoroghi, J. Wu, R. Wang, P. Anand, M. A. Walker. Modelling protagonist goals and desires in first-person narrative //arXiv preprint arXiv:1708.09040. 2017.

17. Tangherlini 2018 — Tangherlini T. R. Toward a generative model of legend: Pizzas, bridges, vaccines, and witches //Humanities, 7, 1, 2018. P. 1–19.

18. Yan et al. 2019 — X. Yan, A. Naik, Y. Jo, C. Rose. Using Functional Schemas to Understand Social Media Narratives // Proceedings of the Second Workshop on Storytelling, 2019. – P. 22-33.

Приложения

Приложение 1. Ссылка на репозиторий с кодом

https://github.com/laidhimonthegreen/prozhito_thesis/

Приложение 2. Источник данных

https://www.dropbox.com/sh/8vfjjt8107sv9r3/AADOVR795MxFovpuGN9PT_JZa?dl=

Приложение 3. Инструкция для разметчиков

“Вам предстоит оценить качество автоматической классификации для N случайно выбранных дневниковых записей.

Данные будут получены в формате таблицы, где:

- первый столбец (note_id) — уникальный ID записи,
- второй столбец (notes) — текст записи,
- третий столбец (PRED) — присвоенный автоматической разметкой тег,
- четвертый столбец (TRUE) — [проставляется вами всегда] 1 или 0 в зависимости от того, верно или неверно был тег определен автоматически
- пятый столбец (TAG) — [проставляется вами при несогласии с тегом] верный, на ваш взгляд, тег
- шестой столбец (TAG-2) — [!необязательно; проставляется относительно редко, при серьезных сомнениях] категория, которую вы тоже рассматривали в качестве верной, но не выбрали

Виды тегов:

— **NAR** (“стандартный нарратив для дневника, перечисление событий”)

Описание повседневных действий. Как правило, в форме перечисления тех или иных событий.

Пример:

“Пятница — днём.

Сегодня я и Шура не пошли в гимназию, потому что у меня и у Шуры болит горло.

18-го папа привёз пуд сахара, масло и муки. Теперь сахар дают по карточкам по 3 фунта на человека. А папа где-то купил без карточек, поэтому и купил пуд.

Эти дни стояли сильные морозы. Было –24. Сегодня потеплело, утром было –13½.

Вчера в гимназии выдавали русский диктант и у меня 2 ошибки; жалко только, что у нас отметок не ставят, а то бы я, наверное, получила 4.

Мать Шепелевой к маме не пришла, потому что её сын заболел. Сейчас должна прийти m-lle.

Я совсем забыла написать очень неприятное для меня событие. 25-го ноября я дала Уткиной книгу «Чёрная птица и орёл снеговых вершин». Прошло две недели, а она мне давала книгу. Дала только 14 декабря. Оказалось, что она её залила чернилами. Мне очень жалко”.

Также включает в себя перечисление событий за более долгий период, чем день (например, за прошедшую неделю).

“Вторник.

Началась третья четверть. Я сейчас учусь хорошо, вызывали меня в последние и ставили только пятёрки и четвёрки. Но вот с алгеброй я прямо не знаю, что мне делать. У нас была недавно контрольная работа. Марьяна мне отмет

ку не поставила, за тетрадь поставила четвёрку. Это алгебра просто поперёк горла у меня застряла.

Сегодня физичка дала мне написать к 5 февраля 1953 года доклад о Ладыгине. Я пошла в школьную библиотеку, там ничего для доклада нет. В Покровскую детскую библиотеку я не хожу. Галя Кочетова обещала спросить книгу в 34 библиотеке, но сегодня она закрыта. Я хотела взять абонемент у Светланы и ехать в Филиал юношества на Красную площадь, да сегодня очень холодно, под вечер обещали минус 30 градусов. Сейчас у меня ужасный насморк и кашель. К вечеру повышается температура. Мне и хочется заболеть и нет. Заболеешь — дома скучно лежать, книг интересных нет, да школу не очень-то хочется пропускать.

Как то у нас с мамой был разговор по душам. Она мне рассказала, что однажды выйдя в коридор она услыхала, или верней подслушала наш разговор с Лилей, а я как раз ходила к Лёвке и рассказывала ей. Таким образом она уже давно знала обо всём.“

— **EPISODE** (“запись об кратком эпизоде”)

Зафиксирован отдельный эпизод из жизни, как правило, во всей записи соблюдены единство времени, места и действия. Подробные описания природы или интерьера тоже относятся к этой категории.

Пример:

“Огромные окна моего класса в художественной мозаике мороза. Словно парчевые ветки мира вырезные листья пальм и глазастые перья павлина раскинулись на стеклах. Ученики тихо пишут изложение по «Красному десанту» Фурманова.

Я просматриваю классные тетради. Оранжевая заря осыпает золотыми блестками верх окон. Низ еще в густо-синем ледяном тумане. Эта картина походит на подводный мир или тропическии лес перед восходом солнца. Напряженные лица учеников. Они вместе с Ганькой метеором мчатся в улагаевской станицу, а по пути вероятно делают ошибку за ошибкой. Ох и достанется же мне проверять их! Еще так безграмотны ученики Дал уже пять диктантов а Ну исчезают туго. За чистоту тетрадей борюсь так, что рву особо неряшливые тетради на глазах их владельцев. И все-же неряшливость еще живет. Скоро и 2 четверть кончается а я не радуюсь ее приближению а беспокоюсь”.

Также может быть зафиксирована последовательность эпизодов (отдельных друг от друга и описанных достаточно подробно):

“Наклюкалась лекарств на травах. Химию не употребляю кроме карворола и в экстренных случаях антибиотиков.

Пришла бабушка моей ученицы по флейте Маши, принесла кефир и молоко,

денег не взяла. Я приготовила подарки для её семьи.

Человек тот прислал мне письмо, что он мне предложение дружить в Агенте не присылал. Я ему это письмо-предложение переслала. Он отнекивается. Странный он всё-таки человек.

Говорила с телефонной приятельницей, она от очень малого общения с людьми стала косноязычной, не может нормально построить фразу. Хорошо, что она это понимает.

Звонил мой приятель, сказал, что вчера умер его друг, актёр театра на Таганке. В интернет эта новость ещё не пришла. Я помолилась о Сергее Подколзине. Мы два раза виделись с ним у моего приятеля. Он был галантен.

Один неприятный тип четыре гадости мне написал, обвиняет он меня в незнании русского языка, мысли у него наглые и злые. Я посмотрела его стихи — очень слабые и неграмотные. Мне пришлось написать ему ироническую рецензию и его же цитату, добавив картинку смешную с выводком цыплят. Ответа не было.
“

— **EMO** (“эмоциональная запись”)
Описание переживаний и чувств человека.

Пример:
“Эта служба все более угнетает меня. Унижение, всегда в ней бывшее, становится все заметнее, очевиднее. Сбежать от нее, из нее, но куда? Вот и ждешь, когда она сама меня исторгнет, как чужеродного, бесполезного ей, несовместимого. Тут даже не механизм службы повинен, а управляющие механизмом. Их выбирают будто нарочно – в раздражение, в оскорбление, в унижение нам, прочим, полагающимся, надеющимся на ум, знания, культуру. Такие – по сути не нужны.

И писать-то про это тошно. И думать – тошно. Будто жалуешься, но– кому? с какой целью? Оправдаться? Но удел избрал сам, сам и покинь, если больше не можешь”.

Также включает в себя:
— *философские записи-рассуждения*
“Всякое живое существо (а может быть, и неживой предмет), даже какая-нибудь инфузория **осуществляет** **себя** **тем** (и живёт), **что** **стремится** **жить**. Это не тавтология, а выражение того факта, что всякая форма жизни, даже самая примитивная существует не благодаря **автоматическим** процессам в ней и вокруг неё, а прежде всего благодаря своей **активности**, стремлению к существованию, к **самосозиданию,**

самобытию. Или благодаря такому стремлению своих родителей-предшественников. Подобное же и с Промыслом, родственным пражизни, у человека. Тайна всякой жизни в **этом** **«ядре»** **прасамосознания**, имеющемуся в каждом индивиде.”

— эмоциональные отзывы на художественные произведения

“«ТРАКТИРЩИЦА»

Спектакль шел ровней, слаженней, выверенней. Первый выход еще не получается, и не пойму почему... Может быть, песенку свою добавить на выход? Что-то не дотянул.

С Марецкой очень приятно играть. Живое общение, и играет она здорово.

Подавляющее большинство отзывов просто великолепные, а на душе почему-то нет полного удовлетворения. Почему бы? Чего-то мне все не хватает, хотя спектакль очень удачен. Грызет что-то... и не дает покоя.

Нет беспредельной, безоговорочной власти над зрительным залом. Где-то близко. Вот-вот, но нет... Очевидно, это «чуть» и характеризует настоящее искусство.”

— оценка/описание конкретных людей или явлений (особенно их моральных сторон)

“У Вали. С Людой мне было бы проще (привыкла уже к ней, и обе одинаково бедно одеты), но общество Люды всегда будит желание внести коррективы в ее привычки, манеры, быт. Валя же чарующе мила, ее тон, стиль, поведение, облик так обаятельны и до того не нуждаются ни в каких исправлениях, что при взгляде на нее сердце восхищенно и с болью сжимается, и из самых тайников души поднимается грустное сожаление — почему не у меня такая. Она подобно своей матери обладает редким умением красиво жить. А мне эта черта в людях, особенно в женщинах, всегда импонировала.

В период своей предыдущей поездки в Л[енингра]д я не раз убеждалась, что этим искусством владеют не многие. Так, напр., совершенно отсутствует бытовая красота в жизни Веры и Вавы, у Соммер и у некоторых других.”

— записи, которые включают в себя описания реальных событий, но в большей степени раскрывают чувства автора по этому поводу

“Вчера умер Гера Копылов. Как раз сегодня я собирался ему звонить. Не успел. Всегда вот так не успеваешь. Он был уже записан на операцию, которая казалась такой рискованной. Когда я был у него два месяца назад, он показался мне не так уж плох, но он всерьез воспринимал возможность смерти, говорил, что хотел бы передать мне свои рукописи, стихи: жалко, если пропадут, вдруг в них что-то есть. Я говорил, что не к спеху. Обещал мне написать про Габая... Что-то густо стало валить вокруг. Возраст, видно, такой.”

— **WORK** (запись о работе”)

Официальное/практически официальное описание рабочего процесса.

Пример:

“Прилетел из Кабула в Душанбе. Выехал в Нурек. Состоялась встреча с моими избирателями. Сделал доклад в клубе строителей о международном положении и состоянии дел в Советском Союзе. Изложил основные задачи энергетиков по выполнению плана развития отрасли в текущей VIII пятилетке. Сформулировал задачи, стоящие перед строителями Нурекской ГЭС. Ознакомился с ходом строительства алюминиевого завода, ведущегося силами Минэнерго. Встретился с руководством Таджикской республики. Решили ряд вопросов, связанных с оказанием помощи”.

Сложные моменты в разметке:

1. Если вы не можете уверенно определить тег записи, поставьте 0 в ячейку с информацией о верности разметки, и запишите в ячейку с определенным вами тегом тег “OTHER”.

2. Если вам кажется, что в целом запись подходит критериям одной из категорий, но в ней встречаются некоторые особенности другой (например, в записи с повествованием о текущих событиях иногда упоминается эмоциональное состояние автора) — выбирайте тот тег, который, на ваш взгляд, является ключевым. При невозможности уверенно определить однозначно ставится тег “OTHER”.

3. “Пограничные” записи:
- “WORK” /”EMO”/ “NAR” — если в записи наряду с описанием рабочих моментов описываются ярко выраженные чувства автора по этому поводу/моменты, не касающиеся работы, ставится тег “NAR”. Если доля эмоций относительно невелика, ставится тег “WORK”. Официальное описание военных действий также относится к тегу “WORK”. Описание размышлений над какой-либо проблемой или задачей относится к тегу “EMO”.

Пример (тег WORK):
“Проредактировала для БСЭ статью Щеголев (автор Благой<com id="148243788314763"/>), дописала своего Шлецера, дала заключение о Шелехове, работала над Шамилем, но не кончила. Текущая работа в БСЭ. То же в библиографической комиссии. То же в месткоме. В библиотеке КА работала опять над Щербатовым и продолжала работу над Шамилем.
(«Красная новь»<com id="148243788314764"/> обратилась ко мне с просьбой написать к юбилею о Марксе — на тему моей книги! Это весьма торжественно — теперь гоняются за темами моего «Литературного оформления» у Маркса, в то время как та же «Красная новь» отказалась в 1924 г. печатать мою статью «„Капитал“ как художественное целое». Тогда отказалась, а теперь просит! Я это предвидела, между прочим, еще в 1924 г.!) Немного редактировала библиографические карточки по истории пролетариата.”

- “EPISODE”/”EMO”/”NAR” — если в записи наряду с описанием моментов из жизни описаны эмоции автора, тег ставится, исходя из того, какие моменты

преобладают.

Пример (тег EMO):
“Желание выходить из болезни, силой воли преодолеть её.

Рассказы Ги де Мопассана восхищают меня.

Когда певица Полина Виардо начинала петь, она околдовывала людей. «Никогда не думал, что он может так сильно любить.... Он говорил только о Полине». Это чей-то рассказ о Тургеневе по радио «Россия».

У Ивана Сергеевича была дочь Полина от крепостной женщины. Тургенев отдал её на воспитание в семью Виардо. Муж Полины Луи стал другом Тургенева.

Полина сказала как-то: «Мы слишком хорошо понимали друг друга, чтобы заботиться о том, что о нас говорят».

Говорила по телефону с Антошей Старчиком, у него очень красивый голос, он поёт в церковном хоре. Он поможет мне с некоторыми проблемами в компьютере, надеюсь. Я заплачу ему.

Я обратила внимание, что сегодня я стала быстро двигаться. Это говорит о том, что энергии стало прибавляться. Ура!”

Пример (тег EPISODE):
“Гуляли с Данутой вдоль озера. Чаша неба, воздух, ветер, источник с хорошей водой. Муж её сказал мне, чтобы я почаще к ним приезжала. Обо мне он сказал Дануте: «Галя красивая, красиво одета, всё знает, умная».

Бабочка ко мне прилетела в комнату. Я хотела осторожно взять её и выпустить, но она спокойно села мне на руку, я поднесла её к окну, бабочка улетела.

У Ани на даче бабочка села рядом со мной на стол. «Покажи свои крылья», — сказала я ей. Она тотчас раскрыла свои крылья, мы ею любовались, она не спешила улететь.

По радио в течение дня диктор говорит: «Поэзия на радио «Россия», объявляют Пушкина. Во мне раздаются строчки: «Поэт, не дорожи любовию народной». Тут же читают именно эти стихи…”

Если сложно определить преобладающий тег, можно поставить тег “OTHER”.”