ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
"НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
"ВЫСШАЯ ШКОЛА ЭКОНОМИКИ"
ФАКУЛЬТЕТ КОМПЬЮТЕРНЫХ НАУК

Казанцев Сергей Станиславович

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Шутки в сторону: машинное обучение и интерпретируемый искусственный интеллект в задачах генерации юмористических текстов

Jokes apart: machine learning and explainable artificial intelligence for humor generation

по направлению подготовки 01.04.02 Прикладная математика и информатика

образовательная программа «Финансовые технологии и анализ данных»

Научный руководитель
Заместитель руководителя департамента,
Профессор ФКН НИУ ВШЭ

_____

В.А. Громов

Студент

_____

С.С. Казанцев

Москва 2025

Аннотация

Дипломная работа посвящена исследованию жанровых особенностей семантических структур в задачах генерации юмористических текстов. Цель исследования состоит в анализе отличий между юмористическим и литературным жанрами на уровне концептуальных представлений, используя методы машинного обучения. В работе применяются методы обработки естественного языка для автоматического извлечения сущностей и связей, построения концептуальных графов, статистический анализ полученных структур и непараметрические критерии для оценки значимости различий. В качестве источников данных использованы корпуса юмористических текстов (анекдоты, краткие шуточные рассказы) и корпус классической художественной литературы.

Основные результаты показывают наличие статистически значимых различий между двумя жанрами по таким характеристикам концептуальных графов, как плотность связей, центральность концептов, средняя длина выделенных субъектов и другим показателям. Наблюдаемые закономерности подтверждают выдвинутую гипотезу о том, что структура концептов существенно варьируется в зависимости от жанра текста. Среди основных ограничений исследования отмечаются неоднородность и разнородность используемых корпусов, а также неточности автоматического извлечения сущностей и построения графов. Полученные выводы могут быть использованы для разработки систем автоматической классификации текстов по жанру, расширения семантического анализа при создании моделей генерации текста, а также повышения качества и интерпретируемости алгоритмов генерации юмористического контента.

Abstract

This thesis investigates the genre-specific characteristics of semantic structures in the context of humorous text generation using machine learning. The objective of the research is to analyze the differences between humorous and literary genres at the level of conceptual representations. To achieve this, methods of natural language processing have been applied for automatic extraction of entities and relations, construction of conceptual graphs, along with statistical analysis of the resulting structures and nonparametric tests to evaluate the significance of observed differences. The data sources include a corpus of humorous texts (jokes, short comedic stories) and a corpus of classical literary texts.

The main results reveal statistically significant distinctions between the two genres in features of the conceptual graphs, such as graph density, centrality of concepts, average length of extracted subjects, and other metrics. The observed patterns confirm the hypothesis that the structure of concepts varies significantly depending on the genre of the text. The primary limitations of the study include the heterogeneity of the corpora used and inaccuracies in the automatic extraction of entities and graph construction. The findings of this work can be used to develop automated text genre classification systems, to extend semantic analysis in text generation models, and to improve the quality and interpretability of algorithms for humor generation. These insights can also inform the design of more effective humor generation systems.

Содержание

Введение

Актуальность исследования. Юмор является одной из сложнейших форм человеческого творчества для генерации. Несмотря на то, что машинное обучение добилось огромных успехов за последние годы, способность алгоритмов понимать и генерировать юмор остаётся достаточно посредственной. Модели могут генерировать шутки на основе миллионов примеров, однако они не способны надёжно оценить, насколько одна шутка “смешна”, а другая – нет.

При работе с творческим языком и в особенности с юмором важна интерпретируемость ИИ. Без реализации прозрачности трудно говорить о глубоком доверии этим моделям и внедрении их в какие-либо ответственные области. В связи с этим возникает потребность определения того, насколько “смешным” является юмор.

В юморе часто бывают задействованы методы вызова несоответствия, какого-то неожиданного смысла, который и создаёт эффект комичности. С точки зрения лингвистики комический эффект возникает в момент, когда сталкивается два несовместимых смысловых сценария в рамках одной шутки. Формальным представлением сценариев в тексте могут послужить цепочки ключевых сущностей – последовательности концептов. В рамках работы можем предположить, что в юмористических текстах будут выделены более необычные, редко сочетаемые последовательности концептов в сравнении с литературными текстами.

Цель работы – подтвердить или опровергнуть гипотезу о существовании статистически значимой характеристики (или множества характеристик), позволяющих отличить юмористические тексты от литературных. Другими словами, цель работы – выявить количественные специфические черты в

последовательностях концептов, которые могут быть присущи только юмористическим или неюмористическим текстам.

Задачи исследования:

1. Провести обзор современных методов обработки естественного язык на русском языке.
2. Изучить теоретические основы лингвистического анализа юмора.
3. Описать выбранные для экспериментов корпуса данныхю
4. Реализовать алгоритм извлечения концептов из текста. Под концептом понимается последовательность сущностей.
5. Реализовать представление последовательностей концептов в виду графов концепции, в которых вершинами являются сущности, а связи между ними отражают их отношения. Из полученных графов выделить признаки, характеризующие структуру концептов для юмора и литературы.
6. Провести анализ различий между корпусами юмористических и литературных текстов с помощью критериев непараметрической статистики, чтобы определить статистическую значимость
7. Подвести итоги эксперимента и сделать вывод о выделенных характеристиках

Объектом исследования являются юмористические тексты и художественные произведения на русском языке. Объектом анализа являются последовательности концептов внутри этих текстов.

Предметом исследования являются статистические характеристики и признаки последовательностей концептов в графовом представлении.

Другими словами, лингвистические и сетевые параметры, которые в совокупности описывают “концептуальный” почерк юмористического текста.

Методы исследования. В работе будут использованы методы компьютерной лингвистики и статистического анализа. Для обработки текстов применяются алгоритмы NLP: распознавание именованных сущностей (ner), токенизация, морфологический анализ, извлечение ключевых сущностей текста. Для построения графов используются методы моделирования текста как сети. Статистический анализ включает в себя вычисление характеристик и применение непараметрических критериев Манна-Уитни и Колмогорова-Смирнова, расчёт непараметрического размера эффекта с помощью дельты Клиффа и проведение многомерного дисперсионного анализа MANOVA.

Научная новизна. Данная работа проводит исследование статистической значимости характеристик последовательностей концепций в юмористических и литературных текстах на русском языке и является одним из первых в этой области. Будет проверена оригинальная гипотеза о наличии статистически значимых отличий в графовом представлении последовательности концептов.

Практическая ценность. Определение статистически значимых характеристик при сравнении шуточных и литературных текстов может помочь ИИ глубже понять природу юмора, а также помочь заложить основы для интерпретируемых алгоритмов генерации шуток.

Глава 1. Машинное понимание текста и основы генерации юмора

В первой главе будут рассмотрены основные теоретические концепции, связанные с анализом юмористических текстов и обработкой естественного языка.

В данной главе рассматриваются основные теоретические концепции, связанные с обработкой естественного языка (Natural Language Processing, NLP) и анализом юмористических текстов. Приводится краткий исторический обзор развития методов NLP – от статистических моделей на основе *n*-грамм до современных подходов к интерпретации смысла через концепты. Также рассматриваются методы извлечения концептов и графового представления текста, статистические методы анализа текстовых корпусов, а завершается глава обзором теоретических моделей юмора (инконгруэнтность, сценарии, структура шутки). Такое изложение позволит сформировать базу для дальнейшего исследования методов машинного обучения и интерпретируемого ИИ в задаче генерации юмора.

1.1. Основы обработки естественного языка и специфика русского NLP

Обработка естественного языка (Natural Language Processing, NLP) – это область искусственного интеллекта и вычислительной лингвистики, изучающая компьютерный анализ и синтез текстов на человеческих языках.

В NLP можно выделить классический пайплайн, состоящих из следующих этапов: токенизация текста – разделение на предложения и слова, лексический анализ (определение у слова части речи и его начальную форму), синтаксический разбор и семантических анализ (определение, интерпретация смысла). Разным языкам характерны разные этапы морфологического и синтаксического анализа в силу разных паттернов и правил языка.

В русском языке можно наблюдать свободный порядок слов в сравнении с многими другими языками. Это компенсируется богатым словоизменением и использованием служебных слов с пунктуацией. В русском языке нередко конечный смысл какого-либо высказывания определяется не порядком слов в предложении, а окончаниями и предлогами. И в этом заключается сложность для обработки естественного языка, ведь модели должны уметь правильно распознавать все грамматические формы слов. Русский язык также богат омонимами, для обработки которых требуется создание отдельный словарей и алгоритмов. Создаются и разрабатываются обширные словари словоформ и правила для разбора слов, которые позволят определить грамматические характерстики и леммы слов.

Как итог, особенности “русского” NLP связано с его флективностью и свободой в построении порядка предложений. Для того, чтобы обрабатывать русский язык, необходим большой перечень инструментов, который будет учитывать сложный синтаксис и морфологию русского языка. И всё еще остается обилие особенностей языка, которое может быть неочевидным для моделей – жаргонизмы, поговорки, неологизмы и т.д.

1.2. Модели n-грамм и концептов: история и современные применения

Одними из первых подходов, возымевшим успех в области компьютерной обработки текста, стали языковые модели, в основе которых лежали цепи Маркова – модели n-грам. Затем было введено понятие энтропии текста – генерации связного текста на основе последовательностей символов и слов. Каждое следующее слово предсказывается на основании n-1 слов. Эти модели вычисляют частоты всех встречающих n-грамм и предсказывают, с какой вероятностью появится следующее слово. Они были в основе первых моделей распознавания речи и машинного перевода, активно применялись в задачах автоподставления следующего слова и исправления опечаток и др.

Достоинство *n*-грамм – простота и независимость от языка: модель обучается напрямую на корпусе и отражает частотные закономерности языка. Однако с ростом *n* экспоненциально растет число возможных последовательностей, многие из которых не встретятся даже в большом корпусе. Проблема «проклятия размерности» ограничивала практическое применение *n*-грамм: для надежной оценки 5-грамм и выше требуются огромные объемы данных или сглаживающие техники. Тем не менее, *n*-граммы до сих пор используются – например, при сжатии текста и оценке перплексии модели или в совмещении с нейросетевыми подходами для интерпретируемости. Современные крупные корпусы (такие как Google Ngrams) предоставляют статистику по миллиардам *n*-грамм, что ценно для лингвистических исследований и приложений NLP.

В то время как *n*-граммные модели оперируют последовательностями слов, альтернативой являются модели концептов – представления текста на уровне абстрактных понятий. Под «концептом» обычно понимается некая сущность или смысловая категория, стоящая за конкретными словами. Простейший пример: слова «автомобиль», «машина», «auto» могут

рассматриваться как один концепт «Автомобиль». Модели концептов имеют свои корни в семантических сетях и онтологиях, разрабатывавшихся в когнитивной науке и ИИ. В 1980-е гг. появились тезаурусы типа WordNet, задающие иерархии понятий, что позволило переходить от слов к *synset*-концептам. Концептуальные модели текста представляют документ как набор или последовательность концептов, а не слов. Исторически это позволяло решать проблему синонимии и вариативности: разные слова с близким значением сводятся к одному узлу-концепту, снижая размерность представления и повышая семантическую связность.

Современные применения моделей концептов разнообразны. В информационном поиске концептуальные индексы помогают расширять запросы: система оперирует понятиями, охватывая синонимы и связанные термины. В тематическом моделировании и рекомендательных системах концепты (например, категории товаров, жанры фильмов) используются для более осмысленной кластеризации, чем просто слова. В генерации текста с юмором (тема нашей работы) концептуальные модели могут связывать отдаленные идеи: шутки часто строятся на неожиданном сопоставлении концептов, и способность модели «понимать» концепты (например, что *«за замок»* можно как пройти, так и сесть) повышает когнитивную интерпретируемость. Одним из направлений развития являются концептуальные языковые модели, комбинирующие статистический подход *n*-грамм с уровнем понятий. В таких моделях текст сначала преобразуется в последовательность концептов (например, путем замены слов на идентификаторы в базе знаний), после чего строится модель аналогично *n*-граммной. Это улучшает обобщающую способность модели и порой снижает зависимость от конкретного словаря или языка.

1.3. Извлечение концептов: подходы, алгоритмы, библиотеки

Переход от слов к концептуальным представлениям требует реализации алгоритма извлечения концептов, то есть автоматическое извлечение именованных сущностей и объектов, упомянутых в тексте. Можно выделить три основных подхода к извлечению сущности: словарные методы, статистические и нейросетевые.

Словарные методы предполагают использование заранее известных словарей концептов. В словарном методе производится распознавание сущностей в тексте и сопоставление их с концептами из базы. Сложные слова же обрабатываются за счёт обилия примеров в базах с семантическим контекстом и словарем значений. Такой подход хорош потому, что является интерпретируемым. Каждый извлеченный концепт будет привязан к конкретному смыслу из описанной базы знаний, вследствие чего его можно описать и связать с другими концептами. Минус данного подхода заключается в ограниченности баз знаний. Постоянно появляются новые слова, и старые слова обретают новые смыслы в определенном контексте, что базы знаний могут не учитывать.

Статистические подходы не требуют готовых онтологий и извлекают концепты исходя из распределения слов в корпусе. Сюда относятся методы ключевых слов и терминологии. Например, алгоритм RAKE (Rapid Automatic Keyword Extraction) находит в тексте ключевые слова и словосочетания на основе частоты и позиции стоп-слов. Он выделяет кандидатов (последовательности слов без стоп-слов) и присваивает им вес, учитывая, насколько часто слова встречаются внутри этих последовательностей и вне их. Аналогично работает алгоритм YAKE и другие. Другой пример – TextRank, графовый алгоритм, примененный к задаче ключевых слов: слова текста связываются ребрами по совместной встречаемости, после чего вычисляется

мера центральности PageRank для узлов-слов; наиболее «важные» слова или фразы и принимаются за ключевые (концепты текста). Графовые методы хороши тем, что учитывают связи между словами: так, даже редкое слово может получить высокий ранг, если оно связано со многими другими частыми словами (что сигнализирует о его тематической значимости). В корпусной лингвистике распространены также статистические критерии значимости терминов: *TF-IDF* (взвешивает частоту слова в документе обратной частотой в корпусе), меры типа *weirdness* (сравнение частот слова в специализированном корпусе и в общем языке) и специализированные метрики для коллокаций – например, *MI (взаимная информация)*, *t-score*, *лог-правдоподобие* и др. Они позволяют автоматически выявлять устойчивые сочетания и специфичную лексику. Статистические методы зачастую выявляют концепты, не заданные явно в словаре, что важно для постоянно меняющихся доменов (новые технические термины, мемы и пр.). Недостаток – полученные кандидаты-концепты могут требовать интерпретации: алгоритм не знает смысла, он лишь находит потенциально значимые словосочетания.

Подходы машинного обучения в извлечении концептов подразумевают обучение моделей на размеченных данных: когда в текстах вручную отмечены концепты (термины, сущности), можно тренировать классификаторы или секвенсорные модели (например, биLSTM или трансформеры) для воспроизведения этой разметки. Классический пример – задача *Named Entity Recognition (NER)*, распознавания именованных сущностей, решается сейчас нейросетевыми моделями, обученными на корпусах с разметкой PER/ORG/LOC и т.д. (имена людей, организаций, географические названия). Такие модели, как BERT**,** RoBERTa, при дообучении на данных по сущностям могут автоматически отмечать упоминания концептов в тексте с высокой точностью. Появляются и универсальные архитектуры, объединяющие несколько задач – выделение сущностей, извлечение отношений между ними и разрешение кореференции – в единой нейросети. Это приближает нас к

автоматическому построению семантических графов из текста. Тем не менее, нейросетевые решения обычно менее интерпретируемы: они выучивают шаблоны, но не дают явной семантической привязки к известному концепту (если не использовать дополняющий слой, привязывающий к базе знаний). Поэтому в контексте интерпретируемого ИИ (важного для нашей темы) часто применяют гибридные подходы: сначала с помощью правил или статистики выделяют кандидатов-концептов, а затем уточняют их классификатором; либо наоборот – нейросеть предлагает кандидатов, а база знаний фильтрует и интерпретирует их.

С точки зрения инструментов, на практике используются сочетания вышеописанных подходов. Для русского языка доступны такие библиотеки, как: Natasha и spaCy с моделями для русского, Tomita-parser от ABBYY (правила и словари). Для терминологических концептов – отдельные разработки: например, алгоритмы извлечения терминов Астраханцева и др., реализованные в виде утилит. Российская Texterra предоставляет API для семантического анализа: на вход подается текст, на выходе – размеченный концептами граф знаний.

Итак, извлечение концептов – ключевой этап на пути к тому, чтобы компьютер «понимал» текст на смысловом уровне. Выделенные понятия служат строительным материалом для графового представления знаний и для генерации осмысленных (а в нашем случае, потенциально смешных) текстов. Ведь, распознав в тексте концепты, система может оперировать ими, комбинировать и искать между ними нестандартные связи – что, как увидим, лежит в основе юмора.

1.4. Представление текста в виде графа: графы слов и сущностей, метрики

Графовое представление текста – это подход, при котором элементы текста (слова, понятия, сущности) рассматриваются как узлы графа, соединенные ребрами в соответствии с определенными отношениями. Такой способ представления позволяет применять теорию графов и сетевой анализ к лингвистическим данным. Существуют разные виды графов текста, два основных – граф слов и граф сущностей.

Граф слов обычно строится на основе лексической близости слов в тексте. Один из распространенных вариантов – *граф соседства слов*: каждое слово в документе является узлом, и проводится ребро между двумя словами, если они встречаются рядом или в пределах фиксированного «окна» (например, 2–3 слов) в тексте. Получается сеть совместной встречаемости (co-occurrence network), отражающая контекстные связи между лексемами. Другой вариант – *граф последовательности*: узлы – это слова, а ребра направлены от каждого слова к следующему за ним в предложении (фактически, визуализация текста как ориентированного графа переходов). В любом случае, граф слов представляет структуру текста не линейно, а через сеть связей, что открывает возможности для анализа. Например, алгоритм TextRank представляет текст как ненаправленный граф слов (узлы – слова, ребро – совместная встречаемость в предложениях) и с помощью итеративной оценки важности узлов (аналог PageRank) выделяет самые значимые вершины – они оказываются ключевыми словами текста. Такого рода графы применяются для *экстрактивного резюмирования*: вычислив важность узлов-слов, затем находят предложения, наиболее покрывающие топовые узлы, и берут их в резюме. Графы слов полезны и для нахождения сообществ слов – кластеризация графа выявляет тематически связанные группы лексики. Кроме того, через анализ центральности можно определять слова, играющие *связующую* роль между темами (высокая промежуточная центральность), что

может быть важно для обнаружения переходов или шутливых каламбуров, связывающих далекие темы.

Граф сущностей или знаний (knowledge graph) – более семантическое представление. В таком графе узлы соответствуют сущностям или концептам (выделенным из текста именам людей, организаций, объектам, либо абстрактным понятиям), а ребра обозначают отношения между ними. Откуда берутся ребра? Один подход – использовать семантические отношения: например, из текста с помощью синтаксического парсинга извлекаются тройки “субъект – предикат – объект” (Subject–Predicate–Object). Каждая такая тройка указывает, что две сущности связаны определенным отношением (предикатом). Например, из предложения “Алексей написал книгу” можно извлечь тройку (Алексей) –[написал]→ (книга). Набор таких троек по всему тексту формирует ориентированный граф знаний, в котором вершины – конкретные сущности (“Алексей”, “книга”), а дуги помечены типом отношения (“написал”). Используя грамматические зависимости, можно достраивать граф связей для любого предложения: современные парсеры выделяют до 30 типов отношений (подлежащее, объект, обстоятельство, владение и пр.). Разумеется, не все эти связи важны – часто выделяют основные: кто с кем действует, кто чем обладает и т.д. В результате граф сущностей представляет, по сути, семантическую сеть текста – подобную тем, что вручную создаются в базах знаний, но полученную автоматически. Такой граф позволяет отвечать на вопросы (поиск пути между узлами может означать объяснение, как факты связаны), выявлять центральных персонажей или объекты (через меры центральности), и даже обнаруживать «структуру повествования». Например, в художественном тексте граф связей между персонажами (кто взаимодействует с кем) дает социальную сеть произведения.

Другой способ построения графа сущностей – граф со-сущностей: узлы – упомянутые сущности, а ребро проводится, если две сущности упоминаются

совместно в каком-то контексте (например, в одном предложении или абзаце). Такой граф часто используется для анализа новостных текстов и биографий, чтобы понять связи: если «Иван Иванов» и «Acme Corp.» часто появляются в одних и тех же статьях, вероятно, между ними есть связь (Иванов – сотрудник Acme Corp. или участник события, связанного с компанией). Графы совместного упоминания позволяют выявлять кластеры связанных сущностей (сообщества) и ключевые *хабы* – сущности, связывающие разные группы (в журналистике это метод для расследования связей лиц и организаций).

Пример метрик для анализа графов текста включает классические меры теории сетей: степень узла (сколько связей у слова или сущности – показатель его частоты или «популярности» в тексте), коэффициент кластеризации (насколько тесно связаны между собой соседи узла – индикатор того, образует ли слово тематический кластер или выступает мостом между кластерами), средняя длина пути между узлами (глубина связи понятий в тексте) и др. Исследования показали, что графы слов, построенные на больших корпусах, обладают свойством *«малого мира»*: очень короткое расстояние между любыми двумя словами при достаточно высокой кластеризации. В частности, для английского языка средняя длина пути в графе соседства слов составляет порядка 2–3, а коэффициент кластеризации ~0.7. Это означает, что лексическая сеть языка высоко связна: через пару промежуточных ассоциаций можно связать практически любые два слова. Такая связность частично объясняет, как в юморе могут сближаться далекие понятия – в семантическом графе языка между ними часто найдется короткая тропинка.

Другой важной метрикой является центральность узла. Различают несколько видов: степень (уже упомянутая простая центральность по числу связей), близость (обратное среднее расстояние до всех других узлов), посредничество (доля кратчайших путей, проходящих через узел). В контексте текста степень может соответствовать частотности термина, близость – обобщенной связанности слова со всеми темами текста, а посредничество –

роли слова-посредника между разными темами. Например, в научной статье слово с высоким посредничеством может быть термином, соединяющим две области знаний. PageRank – известная метрика центральности, изначально примененная в ранжировании веб-страниц, – также находит применение: в TextRank для ключевых слов страницы с высоким PageRank соответствуют словам, часто со-появляющимся с другими важными словами.

Применение графов в анализе текстов предоставляет наглядные показатели связности и структуры информации. В частности, для задачи генерации юмора графовые представления перспективны тем, что позволяют формализовать поиск неожиданных связей между понятиями. Шутка нередко возникает, когда устанавливается редкая, но логически возможная связь между двумя далёкими концептами. Если представить знания в виде графа, такой эффект можно попытаться получить, найдя нетривиальный путь на графе понятий. Например, если граф знаний содержит пути: *«еж – колючий»*, *«кактус – колючий»*, то можно связать ежа и кактус через общий признак и придумать шутливое сравнение. Таким образом, графовое представление и метрики на графах (центральности, сходство путей, плотность связей) – мощный инструмент для разработки и интерпретации алгоритмов генерации юмористических текстов.

1.5. Методы статистического анализа текстов: непараметрические критерии и корпусная лингвистика

Статистический анализ текстов подразумевает применение методов математической статистики для выявления закономерностей и значимых отличий в языковых данных. В корпусной лингвистике, где исследуются большие коллекции текстов (корпусы), часто возникает задача сравнения показателей между разными группами текстов: например, частоты слов в текстах разных авторов, средней длины предложений в разных жанрах, распределения частей речи и т.д. При этом данные далеко не всегда подчиняются нормальному распределению – напротив, частоты слов обычно распределены по закону Ципфа (немного очень частых слов и множество редких). Поэтому для проверки гипотез о различиях применяют преимущественно непараметрические критерии, не требующие предположения о нормальности распределения.

Рассмотрим пример. Допустим, нужно выяснить, отличаются ли по средней длине предложений тексты юмористических рассказов от текстов научных статей. У нас есть две выборки длин предложений (по предложениям из каждого корпуса). Поскольку распределения длины предложений, скорее всего, не нормальны (много коротких, несколько очень длинных), классический t-критерий Стьюдента неприменим без нормализации данных. В такой ситуации используют ранговые критерии. Один из наиболее распространенных – U-критерий Манна–Уитни (эквивалент рангового критерия Уилкоксона для несвязанных выборок). Этот непараметрический тест проверяет, различаются ли медианы двух выборок. Он ранжирует все наблюдения по величине и оценивает, существенно ли элементы одной выборки «превосходят» элементы другой. Если оказывается, что, скажем, предложения в научных текстах почти всегда длиннее, чем в юмористических рассказах, U-критерий покажет статистически значимое различие. Подобным образом, для связанных выборок (например, сравнение стиля автора *до* и *после*

редактуры) используется критерий знаков или критерий знаковых рангов Уилкоксона (для парных наблюдений).

Другой часто используемый метод – критерии согласия для частотных распределений. Критерий χ^2 (хи-квадрат) Пирсона применяется, когда нужно проверить, независимы ли два распределения категорий. Например, предполагается, что в юмористическом корпусе распределение частей речи отличается от новостного: доля междометий и эмоционально окрашенных слов выше. Составляется таблица сопряженности (например, строки – части речи, столбцы – корпус A и корпус B, значения – частоты). Критерий χ^2 проверяет нулевую гипотезу об одинаковости распределений. Если вычисленное значение статистики превышает критическое, гипотеза отвергается – значит, различия значимы. Аналогично, G-критерий (критерий логарифма правдоподобия) используется для тех же целей; он более точен на малых частотах и тоже базируется на сравнивании наблюдаемых частот с ожидаемыми при случайном распределении. Эти критерии широко применяются для выделения ключевых слов корпусов: сравнивая частоту слова в данном корпусе с его частотой в референтном корпусе, можно статистически обосновать, что слово употребляется «чаще, чем ожидалось». Программа WordSmith Tools, к примеру, использует лог-вероятностный коэффициент для обнаружения ключевых слов, характерных для определенного автора или жанра.

Важным направлением является также корреляционный анализ текстовых показателей. Часто возникает вопрос: связаны ли между собой два лингвистических признака? Например, коррелирует ли длина предложения с его комичностью (предположим, что короткие фразы воспринимаются как более шутливые)? Для таких случаев применяют меру ранговой корреляции Спирмена ρ или коэффициент Кендалла τ – они тоже непараметрические и измеряют монотонную связь между переменными, не требуя линейности или нормальности. Эти коэффициенты используются, например, в стилометрии:

проверить, связана ли частота местоимений с эмоциональной тональностью текста и т.д.

В целом, статистические методы в корпусной лингвистике позволяют формализовать выводы о языке. Они дают количественное подтверждение гипотез (например, что в юморе значимо чаще встречаются определенные конструкции). Применение непараметрических критериев особенно оправдано, так как языковые данные часто имеют произвольные распределения и шкалы. Таким образом, исследования в области юмористического текста могут опираться на эти инструменты: проверять, отличается ли статистически речь комиков от обычной разговорной речи, какие слова выбиваются по частоте, есть ли значимые корреляции между использованием тех или иных лексических полей и степенью комического эффекта и т.п. Такие количественные результаты укрепляют лингвистические выводы и позволяют более строго обосновать выбор моделей для генерации шуток (например, выявить, какие признаки текста наиболее сильно ассоциированы с юмором).

1.6. Теоретические подходы к анализу юмористических текстов: инконгруэнтность, сценарии, структура шутки

Юмор на протяжении веков пытались объяснить через различные теории. К классическим относятся теория превосходства, теория облегчения и теория несоответствия (инконгруэнтности). Для анализа именно вербального юмора наиболее плодотворной оказалась теория инконгруэнтности – согласно ей, комический эффект возникает из столкновения несовместимых смыслов или планов повествования в одном тексте. Современные лингвистические модели юмора развивают эту идею с введением понятия «сценариев» или «скриптов».

В 1985 году Виктор Раскин предложил семантическую теорию юмора, сфокусированную на анекдотах и шутках. Основной постулат: текст шутки способен активировать два разных сценария (скрипта), которые находятся в отношении противопоставления или несовместимости друг с другом. Под *сценарием* понимается когнитивная структура знаний о типичной ситуации или объекте. Например, сценарий «похороны» и сценарий «вечеринка» – совершенно разные наборы ожиданий. В удачной шутке присутствуют два плана смысла: первоначальный (ведущий по одному сценарию) и альтернативный (неожиданно раскрывающийся, противоречащий первому). Согласно Раскину, для возникновения юмористического эффекта должны выполняться два условия: в тексте имеется элемент несоответствия, двусмысленности или противоречия, и две интерпретации текста “оппозиционны” в определенном отношении/ Классический пример: шутка «Идет солнце по небу, мать честная» – вызывает комизм, потому что первая часть настраивает на научно-астрономический сценарий, а вторая внезапно переключает на сцену с возгласом испуганной женщины (бытовой сценарий, в котором «Солнце» – вовсе не светило, а зовут так человека или животное). Две разные интерпретации одной фразы сталкиваются, создавая комическое несоответствие. Исследования показывают, что инконгруэнтность сама по себе необходима, но недостаточна для юмора – важна еще фаза разрешения

или объяснимости несоответствия. Если в тексте просто два несвязанных абсурда, юмор может не возникнуть; зато, когда несоответствие разъясняется вторым скрытым смыслом, получается «ага-эффект».

Для нашей работы, посвященной изучению генерации юмористических текстов средствами машинного обучения, эти теоретические идеи предоставляют ключевые ориентиры. Интерпретируемость ИИ здесь помогает: имея явное представление о сценариях (например, через концепты и графы, обсуждавшиеся ранее), модель может отслеживать, что первый сценарий – один набор концептов, второй – другой, и специально вводить противопоставление. Кроме того, понимание структуры шутки (сетап–панчлайн) позволяет формально задать задачу генерации как двухфазный процесс. Теория также подсказывает оценочные критерии: удалось ли модели создать два различных сценария? есть ли между ними осмысленное отношение (противоположность, неожиданная связь)? Таким образом, опора на научные модели юмора обеспечит осмысленность и качество генерированных текстов, отличая их от просто бессмысленных или случайных фраз.

Выводы и результаты по главе

В первой главе был проведен обзор базовых теоретических понятий, необходимых для исследований в области генерации юмористических текстов с помощью методов машинного обучения и интерпретируемого ИИ. Рассмотрены основы NLP: историческое развитие от статистических моделей (*n*-грамм Маркова, работ Шеннона) до современных концептуальных представлений, а также специфические аспекты обработки русского языка (богатая морфология, свободный порядок слов), которые следует учитывать при разработке моделей. Описаны модели *n*-грамм и концептов: статистические *n*-граммы обеспечивают базу для вероятностного предсказания текста, тогда как концепты и семантические сети повышают уровень понимания и интерпретируемости. Далее разобраны подходы к извлечению концептов из текста – от правил и словарей до нейросетевых алгоритмов – и подчеркнута важность этого этапа для перехода от текста к графу знаний.

Было показано, что текст можно представить в виде графа (слов или сущностей), что открывает возможность применять метрики сетевого анализа (центральность, кластеризация, путь и др.) для выявления структуры и ключевых элементов текста. Это особенно полезно для понимания и создания юмора, где зачастую эффект достигается через связи между, казалось бы, далекими концептами – графовая модель облегчает поиск таких связей. Также освещены статистические методы анализа текста: использование непараметрических критериев позволяет подтверждать лингвистические гипотезы на корпусах (например, о специфике языка юмора) строго количественно. Это создает мост между качественными теоретическими наблюдениями и их количественной проверкой.

Наконец, рассмотрены лингвистические и когнитивные теории юмора, центральной из которых является теория инконгруэнтности с концепцией двух

конфликтующих сценариев в шутке. Проанализирована структура шутки (сетап и панчлайн) и условия, необходимые для комического эффекта. Эти теоретические идеи будут служить методологической основой при проектировании системы генерации юмористических текстов: модель должна уметь создавать и распознавать сценарии, вводить несоответствия и выдавать “панчлайн” – все это в рамках интерпретируемого подхода, где решения модели можно объяснить через понятия и связи.

Таким образом, глава заложила теоретический фундамент для дальнейшей работы. Понимание лингвистических основ и моделей представления знаний позволит в следующих главах перейти к обзору современных технологий машинного обучения, применимых для генерации текста, и разработке собственной модели, сочетающей алгоритмы генерации с обеспечением интерпретируемости результатов в задачах юмора.

Глава 2. Практическая часть исследования

2.1. Описание корпусов

В рамках исследования были сформированы два текстовых корпуса: один литературный (неюмористический) и один юмористический. Литературный корпус служит контрольной группой для сравнения с юмористическими текстами. Ниже приводятся характеристики и происхождение каждого корпуса, а также основные этапы предварительной обработки данных (очистки и нормализации текста).

Литературный корпус (неюмористический): включает прозаические тексты, не содержащие явного юмористического содержания. В состав корпуса вошли всевозможные художественные произведения из классической и современной литературы. Источником данных послужили общедоступные электронные библиотеки и корпуса русского языка. Общее число текстов в литературном корпусе – 6429 (суммарно около 92 млн. слов), однако для сравнения с шуточными текстами была взята треть корпуса. Средний объем одного текста – порядка 14,5 тысяч слов. Каждый текст представляет цельное narrativное высказывание без намеренных шуток. Предварительная обработка включала удаление служебной информации (например, номеров глав, иллюстраций, сносок), нормализацию формата (кодировки UTF-8), а также разбиение на отдельные предложения и токены. В ходе предобработки была удалена мета-информация (авторы, рейтинги).

Юмористический корпус 1 (анекдоты): включает короткие анекдоты и шутки, преимущественно в форме устных рассказов и диалогов. Данные получены из открытых интернет-источников, в том числе датасеты с сайта HuggingFace. Корпус содержит порядка 500 тысяч анекдотов и шуток на русском языке (около 11 млн. слов в сумме). Как правило, каждый анекдот –

это короткий текст объемом ~2–5 предложений (средняя длина ~20 слов), зачастую передающий юмористическую ситуацию или диалог с неожиданной концовкой. Тексты анекдотов были разбиты на предложения и токены. Особое внимание уделено очистке от шумов: удалены посторонние символы (эмодзи, нестандартные знаки), исправлены распространенные опечатки.

Сводные количественные характеристики корпусов представлены в таблице 2.1.

Таблица 2.1. Основные характеристики использованных корпусов (объем и средняя длина текстов)

Корпус	Количество текстов	Общий объем (слов)	Среднее слов в тексте
Литературные тексты	2 тыс.	~29 млн.	~14,5 тыс.
Шуточные тексты	500 тыс.	~11 млн.	~22

Предварительная обработка (общая): для корректного выделения семантических связей в литературных текстах важной частью является предобработка текста в корпусах. Тексты были очищены от различных ссылок и неактуальных для произведений примечаний, от лишних знаков препинания, слов и предложений на других языках (французский в классической литературе и английские слова в некоторых анекдотах) и другие фильтрации текста, помогающие уменьшить количество шумов впоследствии.

Таким образом, на выходе получили два очищенных корпуса, в котором каждой строке соответствует индекс и отдельный текст - шутка или произведение, соответственно, готовых к извлечению концептов и построения семантических структур.

2.2. Алгоритм извлечения концептов

Как можно наблюдать в таблице 2.1, юмористические тексты в среднем значительно короче литературных – 20 слов в юморе против десятка тысяч в литературе. Так как далее нами будут сравниваться последовательности характеристик графов концепции, было принято решение о сегментации литературных текстов на примерно сопоставимые с шуточными текстами размеры. В результате корпус литературных текстов был поделен не на произведения, а на сегменты по 2-3 предложения для извлечения концептов.

Чтобы интерпретировать юмористические и литературные тексты, был реализован алгоритм извлечения концептов – ключевых сущностей и различных смысловых единиц текста (персонажи, объекты или важные словосочетания) и связи между ними. Целью извлечения концептов является формирование из текста определенного набора элементов, отражающих абстрагированно, но максимально, насколько это возможно, точно его содержание. В такой форме впоследствии будет удобнее проводить анализ. Алгоритм состоит из следующих этапов:

Выделение именованных сущностей. В первую очередь мы извлекаем из текста именнованные сущности и именные группы. Это могут быть персоны, географические названия, названия магазинов и т.д. Например, из “Антон Павлович ушёл в Пятёрочку” мы извлекаем “антон павлович” и “пятёрочка”.

Разрешение анафоры. Очень часто объекты, которые уже были упомянуты, встречаются дальше в текстах в виде местоимений. Для сопоставления таких местоимений с предыдущими именованными сущностями используется функция solve_anaphora. Если она находит соответствие между местоимением и именованной сущностью – она заменяет его на соответствующий концепт. За счёт этого улучшается связность графа и сохраняется семантика субъекта в случае многократного упоминания.

Формирование отношений «субъект–предикат–объект». В результате для каждого обработанного текста составляется список семантических троек с помощью реализованной функции get_relations. Данная функция ищет предикаты и сопоставляет им ближайшие субъекты и объекты. На выходе мы получаем список отношений (subj, pred, obj) для каждого текста в условиях шуточного корпуса и для набора предложений в условиях литературных текстов.

Чтобы не потерять связи, важно объединять разные forms одних и тех же слов в один узел. Для этого все слова в процессе извлечения концептов лемматизируются и морфологически нормализуются. Для вывода лишних слов, которые не несут в себе значимой семантической информации, фильтруются распространенные стоп-слова и предлоги.

Таким образом, в результате работы алгоритма из обоих корпусов мы получаем по набору концептов – множеству значимых сущностей и связей между ними, представляющих семантическое содержание текста. Данный имеют следующие вид:

```
segment_id,pair_type,subj,pred,obj,union_node,relation_label
lit_47,1,взгляд,на,причина,,nmod
lit_47,1,причина,,эпидемия,,nmod
lit_47,1,эпидемия,,самоубийство,,nmod
lit_47,1,время,,жизнь,,nmod
lit_47,1,черта,,время,,nmod
lit_47,1,сотня,,тысяча,,nmod
lit_47,1,право,с,разветвление,,nmod
lit_47,3,дармоедство,и,дармобытие,род,
lit_47,3,возможность,ни,сила,дармобытия,имевших
```

2.3. Представление текстов в виде графов

Следующим после извлечения ключевых концептов шагом является представление каждого набор концепций в виде ориентированного графа. Для этого использовалась библиотека networkx в сочетании с собственными классами Graph и Edge.

Формирование графовой модели даёт возможность отразить структуру связей между извлеченными концептами внутри текста. Это важно для анализа скрытых смысловых связей и потенциальных механизмов юмора (к примеру, появление неожиданной концепции, связь между несвязанными понятиями).

Процесс построения графов:

- Определение вершин (узлов) графа. Каждый уникальный концепт (subj или obj), лемматизированный и нормализованный, становится вершиной графа. Если два и более концептов встречаются в одном предложении – они соединяются ребрами как связанные смысловой общностью ситуации сущности. Если один и тот же концепт встречается в наборе предложений, то вершины объединяются в единый узел со всеми соответствующими ему ребрами с соседями всех предложений, то есть граф строится с агрегированием повторяющихся сущностей в рамках одного текста
- Построение ребер графа. В каждой триграмме (subj, pred, obj) проводится ребро от субъекта к объекту, а сам предикат сохраняется в виде метки на ребре. То есть ребро показывает смысловое отношение между субъектом и объектом троек
- Сохранение графов. В результате для каждого текста (в нашем случае, каждой отдельной шутки и набора предложений из художественных произведений) получаем набор вершин и связей между ними, из которых строится отдельный граф и сохраняется в формате NetworkX

В результате, для каждого корпуса было построено множество графовых представлений последовательностей концепций. Было построено порядка 350 тыс. графов из шуточных текстов и 250 тыс. графов из сегментов литературных текстов.

Важным моментом графового представления текста является сохранение ключевых характеристик текста - граф позволяет визуально представить сеть семантических связей: множество вершин, соединенных ребрами, дающее возможность наглядно обобщить структуру текста. Граф не придерживается строго порядка слов – он фиксирует факт близости концептов. За счёт этого формируются более глубокие связи.

Уже на данном этапе можно говорить о различиях жанров. В силу того, что литературный текст следует одной сюжетной линии, граф на выходе представляет одну хорошо связанную компоненту. Большинство вершин будет связано общими соседями, т.е. у графов будет достаточно высокая плотность. В свою очередь, шуточные тексты (шутки, анекдоты и т.д.) чаще строятся на инкогруэнтности – отсутствии соответствия между концептами, столкновении двух контекстов, появлении неожиданного смысла. В графе чаще будут встречаться не единая сеть вершин, а кластеры вершин, слабо связанных между собой.

2.4. Выделение признаков и метрик

Представив тексты в виде графов, мы переходим от последовательности слов к взаимосвязанной структуре концептов. Так как проверяемая гипотеза заключается в проверке статистической значимости разницы между последовательностями характеристик шуточных и литературных текстов, ставится задача выделения графовых признаков. Помимо них, однако, я выделил также стилистические признаки на основе триграмм subject-predicate-object. В итоге, мы имеем:

- Стилистические признаки на основе триграмм. Они показывают особенности, которые связаны с формой представления текста:
    - Средняя длина предложения: среднее количество слов в предложении
    - Средняя длина субъектов, предикатов и объектов
    - Количество уникальных субъектов, предикатов и объектов
    - Количество триграмм на один текст
    - Стационарная вероятность π
    - Энтропийная скорость H

- Различные числовые признаки на основе графовых представлений:
    - Количество вершин. Показывает количество уникальный сущностей/объектов в тексте
    - Количество рёбер. Показывает количество предикатный связей, то есть отношений, выделенных в тексте
    - Средняя степень вершины. Показывает среднее количество связей на сущность; отражает насыщенность сущностей связями
    - Плотность. Показывает реализованную долю всех возможных связей
    - Число компонент связи. Показывает, сколькими несвязными частями распадается граф
    - Средний размер компонент. Показывает среднее количество вершин на компоненту, что также даёт представление о фрагментации графа

- Средняя длина кратчайшего пути между вершинами. Считается в неориентированном графе, рассчитывается для крупнейшей компоненты и отражает среднюю дистанцию между сущностями
- Диаметр. Рассчитывает максимальную длину кратчайшего пути между вершинами неориентированного графа. Показывает наибольшее расстояние между сущностями в графе
- Средний коэффициент кластеризации. Рассчитывает среднюю долю замкнутых треугольников вокруг каждой вершины. Характеризует склонность к формированию замкнутых триад
- Глобальный коэффициент кластеризации (транзитивность). Рассчитывает долю замкнутых треугольников от всех связных трёхзвеньев. Отражает замкнутость структуры всего графа
- Усреднённая и максимальная центральность по степени. Отражают распределение связей в графе и центральность самых связанных вершин
- Усреднённая и максимальная центральность по близости. Отражают компактность графа и наличие вершин, которые близки ко всем
- Спектральный радиус. Рассчитывает наибольшее по модулю собственное значение матрицы смежности. Отражает общую связность структуры графа. Чем выше значение – тем крупнее и плотнее граф
- Усреднённая и максимальная центральность посредничества. Отражает “проницаемость” вершин через посредников и значение наибольшего количества коротких связей, проходящих через одну из вершин

Таблица 2.2. Средние значения числовых признаков на основе графовых представлений.

	jokes	literature
num_nodes	4.376966	6.482423
num_edges	2.995979	5.372003
avg_degree	1.263483	1.317796
density	0.312001	0.252831
num_components	1.500477	1.890585
avg_component_size	2.917623	3.085726
avg_shortest_path	1.254396	1.267872
diameter	1.697306	1.722138
avg_clustering	0.040126	0.064763
transitivity	0.048097	0.087724
num_triangles	0.094601	0.463569
avg_degree_centrality	0.624014	0.505670
max_degree_centrality	0.767729	0.653887
avg_closeness_centrality	0.320291	0.260147
max_closeness_centrality	0.602572	0.495509
avg_betweenness_centrality	0.016159	0.011272
max_betweenness_centrality	0.052575	0.040068
leaf_node_count	3.168171	4.232926
spectral_radius	0.017945	0.032566
graph_radius	1.179833	1.195505

Произведём короткую оценку признаков перед тем, как приступать к выявлению статистически значимых признаков:

1. Размер и плотность – в среднем в литературных графах больше узлов (6,48 против 4,38) и рёбер (5,37 против 2,99). При этом, несмотря на бóльшее число узлов, шуточные графы чуть плотнее (.312 vs .253), то есть у относительно небольших шуток связи между элементами чуть более «скучены».
2. Связность и компоненты – число компонент литературе выше (≈1,89 vs ≈1,50), т. е. в литературных графах чаще встречаются изолированные фрагменты. Также чуть больше средний размер компоненты (3,09 vs 2,92), несмотря на большее число малых частей. Средний путь и диаметр практически не различаются – во всех графах «расстояния» между связанными узлами очень малы. Узлы степени – в литературе встречается гораздо больше (4,23 vs 3,17) одиночных компонент,

которые не входят в плотные кластеры.

3. Кластеризация и треугольники – коэффициент кластеризации и транзитивность заметно выше в литературных текстах (0,065 vs 0,040 и 0,088 vs 0,048) – в литературных графах «друзья друзей» чаще оказываются также напрямую связанными. Также и число треугольников на узел гораздо больше у литературы (0,46 vs 0,095), что подтверждает более «тесные» локальные сообщества.
4. Центральность:
    - Степень: шутки имеют чуть более высокий avg_degree_centrality (0,624 vs 0,506) и max_degree_centrality (0,768 vs 0,654) – в шутках отдельные вершины чаще являются «звёздами» с большим количеством связей
    - Близость: шутки и здесь впереди (avg 0,320 vs 0,260; max 0,603 vs 0,496) – средний и максимальный темп «доступности» вершин в шутках выше
    - Посредничество: у шуток больше avg (0,016 vs 0,011) и max (0,053 vs 0,040) – ключевые «мосты» между компонентами внутри графа в шутках играют более выраженную роль
5. Спектральный радиус и радиус графа у литературных графов немного больше (0,0326 vs 0,0179 и 1,20 vs 1,18). Это говорит о чуть более растянутой структуре в глобальном масштабе.

Таким образом, если смотреть на признаки отдельно, шуточные графы имеют более компактную и централизованную структуру с выраженным “звездами”, в то время как литературные графы имеют более разветвленную и модульную структуру: в них меньше централизации и больше изолированных компонент, но локальная кластеризация выше.

В итоге мы формируем две матрицы признаков N×M, где N – количество текстов в соответствующем корпусе, M – рассчитанные признаки.

2.5. Статистический анализ различий между корпусами

Нулевая гипотеза в данной работе формулируется как наличие какой-либо статистически значимой характеристики при сравнении характеристик последовательности концепций шуточных и литературных текстов.

Для проверки гипотезы об отличиях между признаками были расчитаны непараметрические статистические критерии, показывающие наличие либо отсутствие статистической значимости. Анализ проведён на очищенных и отсеченных по 99% перцентилю данных, так как в этот 1% входили очень шумные данные с крайне большим количеством вершин как в литературных графах, так и шуточных. Поскольку сравниваются две независимые выборки по множеству признаков, были выбраны Mann-Whintey U-тест и тест Колмогорова-Смирнова в качестве основных критериев. Для контроля ошибок первого рода при множественных проверках применялась поправка Холма-Бонферрони.

Методы проверки различий:

1. **Mann–Whitney U-тест**
   Данный тест не требователен к нормальности данных, он хорошо работает на шумных данных. В нём для каждого числового признака были сравнены распределения в двух выборках.
2. **Тест Колмогорова–Смирнова (KS-test)**
   Тест Колмогорова-Смирнова характеризует общую форму распределения, включая разницу в хвостах множеств. Он проверяет разницу эмпирических функций распределений признака в обоих множествах.
3. **Поправка Холма–Бонферрони**
   Так как проверяется значительное количество признаков, мы добавляем контроль ошибки первого рода при множественных сравнениях по методу Холма-Бонферрони.

Результаты

После отсечения самых зашумленных данных по количеству вершин и ребер, а также заполнения нулями всех NaN несвязных графов. Можно выделить наиболее статистически значимых признаков:

Таблица 2.3. Статистическая значимость графовых признаков

Признак	p_raw_MW-U	p_Holm_MW-U	значимо?	p_raw_KS	p_Holm_KS	значимо?
avg_closeness_centrality	< 1e–300	< 0.05	Да	1.93e–01	< 0.05	Да
density	< 1e–300	< 0.05	Да	1.88e–01	< 0.05	Да
avg_degree_centrality	< 1e–300	< 0.05	Да	1.88e–01	< 0.05	Да
num_components	< 1e–300	< 0.05	Да	1.95e–01	< 0.05	Да
num_nodes	< 1e–300	< 0.05	Да	1.78e–01	< 0.05	Да
max_closeness_centrality	< 1e–300	< 0.05	Да	1.72e–01	< 0.05	Да
num_edges	< 1e–300	< 0.05	Да	1.59e–01	< 0.05	Да
leaf_node_count	< 1e–300	< 0.05	Да	1.72e–01	< 0.05	Да
diameter	< 1e–300	< 0.05	Да	1.95e–01	< 0.05	Да
avg_shortest_path	< 1e–300	< 0.05	Да	1.75e–02	< 0.05	Да
graph_radius	< 1e–300	< 0.05	Да	7.94e–03	< 0.05	Да

Примечание: p_raw < 1e–300 показывает, что фактические p-значения значительно меньше машинного эпсилон.

Помимо проверки статистической значимости при помощи Cliff’s delta была оценена практическая величина отличий. Данный непараметрический размер эффекта показывает, насколько сильно различие между двумя множествами:

$\delta \in [-1,1]$,

$|\delta| < 0.147$ – незначительный эффект,

$0.147 \leq |\delta| < 0.33$ – малый эффект,

$0.33 \leq |\delta| < 0.474$ – средний,

$|\delta| \geq 0.474$ – большой.

Таблица 2.4. Результаты по дельте Клиффа

Признак	δ (Cliff’s Delta)	Размер эффекта
avg_closeness_centrality	–0.22	small
density	–0.22	small
avg_degree_centrality	–0.22	small
num_nodes	+0.22	small
num_components	+0.21	small
max_closeness_centrality	–0.20	small
num_edges	+0.20	small
leaf_node_count	+0.20	small
diameter	–0.20	small
avg_shortest_path	–0.19	small
graph_radius	–0.19	small
all other metrics	≈ 0	negligible

- Отрицательные δ говорят о том, что показатели выше в юморе.

- Положительные δ сигнализируют, что больше значения по этим метрикам в литературе

Все 11 ключевых признаков показали малый размер эффекта равный $0.15 \lesssim |\delta| \lesssim 0.22$. Можно сказать, что присутствует хоть и умеренное, но устойчивое различие между жанрами. Остальные признаки показали незначительный эффект. И хоть статистически они также значимы, однаки практическая значимость крайне мала.

В дополнение к предыдущим непараметрическим проверкам был проведен многомерный дисперсионный анализ MANOVA. Он позволяет проверить сразу весь вектор ключевых признаков на предмет различий между двумя корпусами. В качестве значимых признаков были выбраны следующие показатели: avg_closeness_centrality, density, avg_degree_centrality, num_components, max_closeness_centrality, leaf_node_count, diameter, avg_shortest_path и graph_radius.

Результаты MANOVA:

- **Wilks' $\lambda = 0.9568$**, $F(9, 572\,279) = 2867.79$, $p < 0.0001$
- **Pillai's trace = 0.0432**, $F(9, 572\,279) = 2867.79$, $p < 0.0001$
- **Hotelling–Lawley trace = 0.0451**, $F(9, 572\,279) = 2867.79$, $p < 0.0001$
- **Roy's greatest root = 0.0451**, $F(9, 572\,279) = 2867.79$, $p < 0.0001$

$\lambda = 0.9568$ означает, что ≈ 95.7 % вариации остаётся внутри групп, а 4.3 % объясняется различием между группами вместе с $p < 0.0001$говорит о том, что средние векторы признаков «шутка» и «литература» статистически различаются.

Это дает нам с исчерпывающей статистической уверенностью подтвердить нулевую гипотезу о наличии статистически значимых характеристик, показывающих разницу двух выборок.

Общий вывод

Mann–Whitney U и KS-тест с поправкой Холма–Бонферрони показали статистическую значимость во всех выделенных признаках. Лишь несколько из них были менее значимыми других. Можно утверждать, что практически по всем графовым метрикам литература и юмор различаются ($p < 0.05$ после коррекции).

Также был рассчитан практический размер эффекта дельта Клиффа. Он подтвердил, что различия хоть и малые, но устойчивые. Можно утверждать, что:

- Компактность и центральность выше в шутках
- Разветвлённость и масштаб выше в литературе

Многомерная проверка (MANOVA) на девяти ключевых признаках также дала статистически значимый результат: Wilks' $\lambda = 0.9568$, $F(9, 572\ 279) = 2867.79$, $p < 0.0001$ (Pillai's trace = 0.0432, аналогично $p < 0.0001$). Это говорит о том, что литературные и юмористические тексты различаются не только по отдельным метрикам, но и в качестве многомерных объектов.

Все эти измерения подтверждают, что юмористические тексты представляют из себя более компактные и высокоцентрализованные графы концепции, в то время как литературные графы являются более крупными и разветвленными. Эти характеристики могут служить признаком для автоматического определения жанра и впоследствии послужить одним из параметров при генерации юмора.

Заключение

В настоящей работе были проанализированы корпуса юмористических и литературных текстов с точки зрения последовательностей концептов (n-грамм сущностей) и соответствующих им графовых моделей. Для каждого текста строится семантический граф, где вершины отвечают найденным концептам или словам, а ребра отражают факт совместного употребления концептов в пределах предложений или повествовательных единиц. По предложенной графовой модели вычисляются разнообразные количественные характеристики: веса вершин и рёбер, степени узлов, плотность сети и т.д.. Данный подход предоставляет дополнительные возможности в сравнительном анализе текстов, выходящие за рамки традиционных частотных методов. В результате исследования был получен свод статистических данных по ключевым параметрам каждого корпуса, что позволило объективно сравнить структуры юмористических и художественных текстов.

Проверка исходной гипотезы о наличии статистически значимых различий между двумя типами текстовых корпусов показала подтверждение ожидаемых выводов. Проведённый сопоставительный анализ выявил, что по ряду рассматриваемых признаков корпуса заметно различаются. Имеются свидетельства того, что между юмористическими и не юмористическими текстами (например, литературными) действительно существуют статистически значимые отличия в характеристиках текста. Наши результаты согласуются с общими наблюдениями в лингвистике о различиях жанров: смешные тексты демонстрируют иные паттерны использования концептов, чем традиционные художественные произведения.

В частности, выделены ключевые признаки, по которым обнаружены существенные различия. Одним из таких показателей является плотность графа (отношение числа фактических связей к максимально возможному), которая у юмористических текстов оказалась статистически меньше – это

отражает более фрагментарную, «разбросанную» структуру повествования по сравнению с литературными текстами. Аналогично, существенно различаются меры центральности вершин (например, средняя степень узла или центральность сближения): литературные графы чаще содержат ярко выраженные центральные узлы (главные темы или персонажи), в то время как юмористические тексты характеризуются более равномерно распределённой сетью концептов. Также зафиксированы отличия в длине субъектов (средней длине цепочек сущностей или средним объёмах именованных объектов): у юмористических текстов средняя длина таких цепочек статистически меньше, чем у художественных произведений, что указывает на более краткую и динамичную подачу информации. Помимо этого, нами отмечены различия в коэффициенте кластеризации и усреднённой длине кратчайших путей в графах, свидетельствующих о более низкой взаимосвязности концептов в юморе. Во всех перечисленных признаках (плотность, центральность, длина сущностей и др.) различия оказались статистически значимыми, что подтверждает обоснованность исходной гипотезы.

Наряду с выявленными результатами необходимо учитывать ряд ограничений проведённого исследования. Во-первых, корпуса текстов остаются достаточно неоднородными: в выборку входят произведения разных авторов, периодов и поджанров, что усложняет выводы и может вносить смазывающую вариативность. Во-вторых, анализ основан на автоматическом извлечении сущностей (Named Entity Recognition) и концептов, что не лишено ошибок. Погрешности распознавания (особенно в присутствии аллюзий, игр слов или просторечной лексики) могли влиять на точность полученных n-грамм и, соответственно, на графовую структуру. Также стоит отметить ограничения по объёму выборок: для статистически надёжных выводов желательно иметь большую и сбалансированную базу текстов обоих типов. Наконец, в работе упрощённо предполагается, что всем выделенным

концептам приписывается равнозначный вес, хотя в дальнейшем можно учесть их семантическую и информационную значимость при анализе.

Перспективы дальнейшего развития темы включают несколько направлений. Создание размеченных корпусов юмористических и художественных текстов (с высококачественной аннотацией сущностей и семантических ролей) позволит обучать и проверять более совершенные алгоритмы выделения концептов, сводя к минимуму текущую погрешность. Применение нейросетевых языковых моделей и методов глубокого обучения (например, контекстуальных эмбеддингов) может повысить точность идентификации семантических связей в тексте и помочь выявить более тонкие различия между жанрами. Расширение выборки за счёт включения дополнительных авторов, источников или других жанров (рекламные тексты, публицистика и т.д.) повысит обобщаемость результатов. Интересно также исследовать расширенные семантические сети, учитывающие не только последовательность сущностей, но и их категориальную и онтологическую информацию. В перспективе описанные подходы могут применяться в задачах автоматической классификации жанра текста, генерации юмористических фрагментов и в других областях цифровой лингвистики, что в итоге углубит понимание лингвистических маркеров юмора и литературного стиля.

Таким образом, результаты работы подтвердили наличие стилистических и структурных отличий между корпусами юмористических и литературных текстов. Выявленные закономерности подкрепляют идею о том, что различные жанры обладают отличительными сетевыми характеристиками и «рецептурой» использования концептов. Будущие исследования, учитывающие указанные ограничения и применяющие более совершенные инструментарии, позволят уточнить эти выводы и расширить их применение в практических задачах лингвистического анализа.

Список литературы

1. **Большакова Е. И., Воронцов К. В., Ефремова Н. Э., Клышинский Э. С., Лукашевич Н. В., Сапин А. С.** Автоматическая обработка текстов на естественном языке и анализ данных. – М.: Изд. дом Высшей школы экономики, 2017. – 268 с.
2. **Amin M., Burghardt M.** A Survey on Approaches to Computational Humor Generation // *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. – 2020. – P. 29–41.
3. **Attardo S.** Linguistic Theories of Humor. – Berlin; New York: Mouton de Gruyter, 1994. – 426 p.
4. **Chen Y., Yuan Y., Liu P., Liu D., Guan Q.** Talk Funny! A Large-Scale Humor Response Dataset with Chain-of-Humor Interpretation // *Proceedings of the AAAI Conference on Artificial Intelligence*. – 2024. – P. 17826–17834.
5. **Hempelmann C. F., Raskin V., Triezenberg K. E.** Computer, Tell Me a Joke... but Please Make it Funny: Computational Humor with Ontological Semantics // *Proceedings of the 19th International FLAIRS Conference*. – 2006. – P. 746–751.
6. **Molnar C.** Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. – Lulu.com, 2020. – 318 p.
7. **Petrović S., Matthews D.** Unsupervised joke generation from big data // *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*. – Sofia, 2013. – P. 228–232.
8. **Zhang H., Liu D., Lv J., Luo C.** Let’s be Humorous: Knowledge Enhanced Humor Generation // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Student Research Workshop)*. – 2020. – P. 156–161.