Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ _____ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ_____
КАФЕДРА _____СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ__

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

Исследование распределения слов в текстах на разных языках и классификация текстов

Студент ___ИУ5И-32М___ _____ _____Се Цзявэнь_____
(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель _____ _____В.И. Терехов__
(Подпись, дата) (И.О.Фамилия)

Консультант _____ _____Канев А.И._____
(Подпись, дата) (И.О.Фамилия)

20 24 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ
Заведующий кафедрой ____ИУ5___
(Индекс)
_____ ____В.И. Терехов__
(И.О.Фамилия)
« _____ » _____ 20 24__ г.

З А Д А Н И Е

на выполнение научно-исследовательской работы

по теме _____Исследование распределения слов в текстах на разных языках и классификация текстов_____
Студент группы __ИУ5И-32М_____Се Цзявэнь_____
(Фамилия, имя, отчество)
Направленность НИР (учебная, исследовательская, практическая, производственная, др.)
_____Исследовательская_____
Источник тематики (кафедра, предприятие, НИР) _____КАФЕДРА_____
График выполнения НИР: 25% к _6_ нед., 50% к _5_ нед., 75% к 12 _ нед., 100% к 16__ нед.
Техническое задание *__В данном исследовании рассматриваются методы предварительной обработки текста и их применение в анализе текстов. Мы выбрали подходящие инструменты и технологии, собрали и обработали крупномасштабные языковые данные, провели разбиение на слова и удаление стоп-слов, а также проанализировали частоту и распределение слов. Для лучшего понимания структуры текста было проведено дальнейшее исследование влияния методов сегментации китайского текста на классификацию текстов. Мы использовали алгоритм TF-IDF для извлечения признаков и сравнили результаты классификации текстов с использованием различных методов сегментации с помощью метода опорных векторов (SVM)._____*
Оформление научно-исследовательской работы:
Расчетно-пояснительная записка на _____ листах формата А4.
Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)
_____
Дата выдачи задания « _19 » ____декабрь__ 2024_ г.
Руководитель НИР _____ _____В.И. Терехов__
Студент группы ИУ5И-32М _____ _Се Цзявэнь____
Консультант _____ Канев А.И__

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

СОДЕРЖАНИЕ

Аннотация

С увеличением межкультурного обмена литература переводится на множество языков, что делает сравнительное изучение различных версий произведений актуальным. Одним из таких произведений является роман «Приключения Оливера Твиста», который существует на английском, русском и китайском языках. Различия в лексике и частоте употребления слов влияют на восприятие и сложность текста. Целью данного исследования является анализ частотности слов в трёх версиях романа с использованием статистических методов, таких как формулы Шеннона и Хартли, а также изучение влияния методов сегментации китайских слов на классификацию текстов.

В ходе исследования были выявлены значительные различия в лексическом составе и частоте употребления слов в английской, русской и китайской версиях романа. Английская версия продемонстрировала стабильность в частоте слов, в то время как русская версия показала большую изменчивость в использовании лексики. Китайская версия, в свою очередь, характеризуется богатым лексическим выбором: при использовании полного словаря энтропия информации оказалась высокой, тогда как при применении одиночных символов энтропия была значительно ниже. Дополнительно было отмечено, что текст в разных главах имеет различную информационную сложность, что указывает на вариативность восприятия текста в зависимости от контекста.

Также было проведено исследование влияния методов сегментации китайских слов на классификацию литературных произведений. Для этого были собраны различные наборы данных, включая романы и биографии, и использован алгоритм TF-IDF для извлечения признаков. Сравнив сегментацию на основе символов и многосложных слов, было установлено, что сегментация на основе многосложных фраз превосходит сегментацию по символам по точности, как для больших, так и для малых наборов данных. Это подчеркивает важность выбора метода сегментации для повышения точности классификации китайских текстов.

Результаты исследования имеют практическое значение для лингвистики, теории перевода, а также для культурных исследований и разработки программного обеспечения для перевода. Статистический анализ помогает лучше понять, как различные языки передают информацию и как методы сегментации влияют на обработку текстов. Это может способствовать созданию более эффективных методов перевода и адаптации текстов, а также улучшению классификации литературных произведений.

Ключевые слова

Частотный анализ слов; Энтропия информации; Статистика глав; Текстовый анализ; NLP; Классификация текста; китайская литература; сегментация слов;TF-IDF.

Введение

Язык является важным инструментом человеческого общения, а тексты на разных языках представляют собой богатое и разнообразное распределение слов. Эта область исследований тесно связана с количественной лингвистикой, которая использует данные и количественные методы для анализа языковых явлений. Одним из ключевых направлений количественной лингвистики является изучение частоты слов. С развитием интернета и информационных технологий, огромное количество текстовых данных требует разработки эффективных методов извлечения информации и классификации текста. Классификация текста является важной задачей в обработке естественного языка (NLP), находящей применение в таких областях, как категоризация новостей, анализ настроений и обнаружение спама.

В данной статье основное внимание уделяется произведению «Приключения Оливера Твиста» и приводится подробная статистика частотности слов в его английской, русской и китайской версиях. Мы понимаем, что каждый язык имеет свои уникальные выражения и словарные характеристики. Благодаря таким исследованиям можно интуитивно увидеть различия в использовании слов на разных языках при интерпретации одной и той же истории. В исследовании мы используем формулы Шеннона и Хартли для более глубокого анализа информации, заложенной в распределении слов. Также проводится нормализованная статистика слов или отдельных слов в каждой

главе книги, а также визуализируются данные для наглядного отображения тенденций и закономерностей распределения слов по всему тексту.

Кроме того, в обработке китайских текстов возникают уникальные трудности, связанные с особенностями языка, особенно в отношении методов сегментации, которые существенно влияют на результаты классификации. Эти трудности подчеркивают важность применения современных методов анализа в лингвистике и информатике, способных справляться с различиями между языками и культурами.

Описание предметной области

Исходная информация

Приключения Оливера Твиста (Oliver Twist [1]) — классический роман 19 века, написанный британским писателем Чарльзом Диккенсом, впервые опубликованный в 1837–1839 годах. Это произведение не только занимает важное место в истории литературы, но и пользуется популярностью благодаря глубокому описанию викторианского лондонского общества и глубоким размышлениям о жизни бедных детей. В романе рассказывается о сироте Оливере Твисте, который сбежал из бедности и жестокого обращения и в конечном итоге искал любовь и счастье. Благодаря яркому изображению ролей и богатому социальному контексту Диккенс не только раскрывает социальную несправедливость, но и демонстрирует дух личной настойчивости.

Глобальное влияние «Приключения Оливера Твиста» привело к его переводу на несколько языков, что позволило произведению преодолеть культурные и языковые барьеры и охватить читателей в разных странах и регионах. Перевод включает не только прямое преобразование языка, но и адаптацию культурного и социального контекста [2]. Таким образом, в различных языковых версиях «Приключения Оливера Твиста» могут быть значительные различия с точки зрения выражения, культурной адаптации и читательской приемлемости. Эти различия не только отражают особенности

языка, но и могут пролить свет на понимание и воспроизведение оригинала в процессе перевода.

В процессе перевода переводчикам часто нужно найти баланс между лояльностью к оригиналу и адаптацией к целевому языку. Достижение этого баланса может привести к значительным различиям между языковыми версиями с точки зрения использования слов, структуры предложений, культурных элементов и эмоционального выражения [3]. Например, китайский перевод может адаптировать некоторые социальные контексты или детали ролей, чтобы быть ближе к культурному восприятию китайского читателя, в то время как русский перевод может сохранить некоторые конкретные культурные детали, чтобы сохранить оригинальный социальный контекст.

Углубленный анализ текстов различных языковых версий «Приключения Оливера Твиста» помогает выявить конкретные проявления этих стратегий перевода и способов обработки. Этот анализ не только помогает понять различные интерпретации оригинального контента в различных культурных контекстах, но и выявляет механизмы трансформации языка и культуры в процессе перевода. Анализируя частоту слов, энтропию информации и структуру глав, исследование направлено на выявление языковых особенностей и текстовых различий, связанных с переводом в различных языковых версиях, что дает новые идеи и ссылки для исследований перевода.

Связанные теории в области анализа текста

Область текстового анализа претерпела значительную эволюцию от традиционных статистических методов до современных технологий глубокого обучения. Ранние методы анализа текста, такие как модель "мешок слов" (BoW) [4], относятся к 1950-м годам. Эта модель предлагает простой и интуитивно понятный способ обработки текста, представляя его как распределение частоты слов, игнорируя последовательность слов и грамматическую структуру. Хотя модель "пакет слов" базируется на простых принципах обработки текста, она заложила важную основу для более поздних методов анализа текста и сыграла ключевую роль в ранних задачах обработки естественного языка (NLP).

Векторная пространственная модель (VSM) [5], предложенная Джерардом Салтоном в 1960-х годах, решает проблему разреженности слов в традиционных моделях, представляя документы и запросы в виде векторов и вычисляя их сходство. VSM добилась значительных успехов в области поиска информации и заложила теоретическую основу для последующих моделей, таких как потенциальный семантический анализ (LSA) [6]. LSA была предложена Стюартом Дерманом и Трояном в 1990 году с целью захвата скрытой семантической структуры путем уменьшения семантики слов и документов.

Метод TF-IDF (Term Frequency - Inverse Document Frequency) [7] был предложен в 1972 году ученым-компьютерщиком Карен Спарк Джонс. Этот метод сочетает частоту слов (TF) и частоту обратного документа (IDF) и предназначен для измерения значения слов в конкретном документе, что

повышает точность поиска информации и анализа текста. TF-IDF быстро стал стандартной технологией для поиска текста и информации, играя ключевую роль в поисковых системах и системах рекомендаций и широко использовался в последующих моделях.

С появлением векторных моделей слов, таких как Word2Vec [8], технология анализа текста вступила в новую фазу. Word2Vec был предложен Миколовым и другими в 2013 году для более точного выражения семантических отношений между словами, отображая их в непрерывном векторном пространстве. Этот технологический прорыв усложнил обработку текстовых данных и способствовал развитию технологий глубокого обучения, таких как предварительно обученные языковые модели, например, BERT [9] и GPT [10], которые значительно повысили точность и эффективность анализа текста, захватывая последовательность слов и контекстную информацию.

Материалы и методы

Частотный анализ слов

Словарный анализ [11] - это статистический метод анализа текста, предназначенный для раскрытия основного содержания, темы и характеристик текста путем вычисления частоты появления слов в тексте. Внедрение терминологического анализа требует следующих шагов:

Предварительная обработка текста:

Подразделение (Tokenization) [12]: Разделить текст на отдельные слова или подклассы. В разных языках метод словосочетания может отличаться. Для английских текстов мы выбрали библиотеку NLTK [13]. NLTK (Natural Language Toolkit) — это Python-библиотека, предназначенная для обработки естественного языка в области искусственного интеллекта. Она предоставляет удобный интерфейс для обработки текстовых данных, включая классификацию, аннотирование, синтаксический анализ, семантическое выведение и извлечение информации из текста. Для обработки русских текстов мы выбрали spaCy [14], поскольку spaCy предлагает высококачественные предобученные модели для русского языка (например, ru_core_news_sm, ru_core_news_md и ru_core_news_lg). Эти модели были обучены на крупных корпусах текстов и поддерживают такие функции, как сегментация слов, определение частей речи, распознавание именованных сущностей и другие, что позволяет эффективно обрабатывать русские тексты.

Однако, поскольку в китайских текстах отсутствуют явные границы слов, размытость границ слов требует от системы токенизации использования контекстной информации для определения границ слов. Другим вызовом является частое появление новых слов и разговорных выражений в китайском языке, таких как интернет-жаргон и модные выражения. Эти новые слова и изменяющиеся языковые формы часто не включены в традиционные словари, что требует динамического обновления словаря и обучения моделей для

адаптации к этим изменениям. В рамках данного исследования мы выбрали библиотеку Jieba [15], которая на данный момент является одним из наиболее эффективных инструментов для китайской токенизации на Python. Библиотека Jieba использует китайский словарь для определения вероятности связи между иероглифами, формируя группы слов на основе высокой вероятности связи между иероглифами. В дополнение к токенизации, пользователи могут добавлять свои собственные словосочетания. Для реализации эффективного сканирования графа слов используется префиксный словарь, который генерирует направленный ациклический граф (DAG) всех возможных комбинаций слов в предложении. Затем применяется динамическое программирование для нахождения пути максимальной вероятности, чтобы выявить максимальную комбинацию сегментации на основе частоты слов. Для не зарегистрированных слов используется модель скрытых марковских моделей (HMM) и алгоритм Витерби. Библиотека Jieba поддерживает четыре режима токенизации: точный режим, полный режим, режим поисковой системы и режим Paddle. Она также поддерживает токенизацию традиционного китайского языка и пользовательские словари. Конкретно, точный режим разделяет предложение на наиболее точные сегменты, что подходит для текстового анализа; полный режим сканирует все возможные слова в предложении, обеспечивая высокую скорость, но не решает проблемы неоднозначности; режим поисковой системы является расширением точного режима и повторно делит длинные слова для повышения полноты поиска, что делает его подходящим для поисковых систем. Поэтому в нашем исследовании мы выбрали точный режим.

Восстановление формы слова (Lemmatization) [16]: Восстановление словаря до его основной формы или корня, например, восстановление « учебный, учебная, учебное, учебные» до «учебный». Это помогает стандартизировать терминологию и уменьшает влияние деформации одного и того же термина на статистические результаты.

Удалить деактивированные слова: удалить слова, которые распространены в анализе, но мало влияют на результат (например, "да", "да" и т.д.).

Стандартизированная лексика: унифицированные форматы словаря, такие как перевод всех слов в нижний регистр, удаление пунктуации и т.д. для обеспечения статистической точности.

Расчет частоты слов:

1. Количество слов: Статистика количества случаев появления каждого слова в тексте.

2. Частотная сортировка: сортировка в зависимости от частоты появления словаря, как правило, с высокочастотными словарями, перечисленными впереди, что облегчает дальнейший анализ. В этом исследовании мы выбрали первые 50 слов в качестве объекта исследования.

Метод расчета энтропии информации

Энтропия информации (энтропия Шеннона), возникшая в первой количественной теории связи и передачи информации [17,18], - это мера неопределенности системы (в статистической физике или теории информации),

в частности, непредсказуемости появления любого символа в первичном алфавите. В последнем случае энтропия численно равна количеству информации, приходящейся на один символ передаваемого сообщения при отсутствии потерь информации.

При отсутствии потерь информации информационная бинарная энтропия рассчитывается по формуле Хартли. Формула Хартли или хартлиевское количество информации или мера Хартли — логарифмическая мера информации, которая определяет количество информации, содержащееся в сообщении.

$$\mathrm{I} = K \log_2 N$$

Где:

N — количество символов в используемом алфавите (мощность алфавита), **K** — длина сообщения (количество символов в сообщении), **I** — количество информации в сообщении в битах.

Формула была предложена Ральфом Хартли в 1928 году как один из научных подходов к оценке сообщений.

Для случайной величины x, принимающей n независимых случайных значений x_i с вероятностями p_i (i=1,..n), формула Хартли переходит в формулу Шеннона:

$$H(x) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Здесь:

(- $\log_2 p_i$)означает измеряемое в битах количество информации, содержащейся в том событии, что случайная величина приняла значение x_i

$H(x)$ — количество информации, которое в среднем приходится на одно событие

Нормализация — это процесс стандартизации данных, позволяющий сделать данные разной величины сопоставимыми. В исследованиях распределения слов нормализация может преобразовать частоту слов в относительную частоту или преобразовать тексты разной длины в одинаковую длину для сравнения и анализа.

При изучении распределения слов в текстах на разных языках информативность текста можно оценить путем расчета энтропии Шеннона. Обработка нормализации может сделать данные текстов разных языков сопоставимыми, тем самым более точно сравнивая характеристики распределения слов между ними.

Статистика и визуализация глав

В этом исследовании статистика глав является центральным звеном в понимании структуры и содержания текста. Этот процесс охватывает такие аспекты, как количество слов, количество символов, частотный анализ слов и расчет энтропии информации. Глоссарий подсчитывается путем дробной обработки текста, подсчитывая общее количество слов в каждой главе для оценки содержания глав. Счет символов вычисляет общее количество символов главы после удаления пунктуации и специальных символов, чтобы отразить

количество чтения и плотность текста главы. Анализ частоты слов призван выявить ключевые термины и темы в главах и обеспечить поддержку данных для дальнейшего обсуждения тем путем вычисления частоты слов.

В изуализация данных является жизненно важным инструментом для эффективной демонстрации этих статистических результатов. Штриховые диаграммы предназначены для отображения словарного запаса и количества символов в каждой главе и позволяют четко отображать различия в данных между главами. Сломанные диаграммы используются для отображения результатов вычислений энтропии информации и визуального отражения тенденций изменения плотности информации. Более высокая энтропия информации обычно указывает на большую сложность и объем информации в разделе, в то время как более низкая энтропия информации может указывать на концентрацию или повторение содержания раздела.

Экспериментальное проектирование и сбор данных 1

Выбор языковой версии

Для данного исследования выбраны следующие версии текста " Приключения Оливера Твиста" (оригинальное название "Oliver Twist") на различных языках: английская версия издательства Wordsworth Editions [1], китайская версия от издательства Yilin Publishing House [19], а также русская версия издательства "ХудЛит" [20]. Эти материалы будут использоваться для анализа текста на разных языках.

Определение сферы анализа

Выберите 53 главы книги для анализа, чтобы убедиться, что они охватывают основную часть романа, а текст обрабатывается сегментами для облегчения анализа.

Стратегия словосочетания

Для английского и русского языков методы сегментации достаточно ясны, так как они основаны на пробелах между словами. В исследовании сегментации китайского текста мы рассмотрели три различных подхода, чтобы справиться со сложностью китайского языка. Первый подход предполагает сохранение полных слов, состоящих из одного или нескольких символов, чтобы обеспечить точность сегментации и полное выражение смысла слов. Например, «我喜欢机器学习»(Мне нравится машинное обучение) будет сегментирован как «我»(Мне), «喜欢»(нравится) , «机器学习»(машинное обучение). Второй подход включает сегментацию по отдельным символам, где каждый символ обрабатывается как отдельная единица, например, «我喜欢机器学习» (Мне нравится машинное обучение) будет сегментирован как «我»(Мне), «喜», «欢» (нравится) , «机», «器», «学», «习»(машинное обучение). Третий подход сосредоточен на сохранении слов, состоящих из нескольких символов, игнорируя одиночные символы, и извлекает только многосимвольные слова. Например, «我学习机器学习» (Мне нравится машинное обучение) будет сегментирован и сохранен как «喜欢»(нравится), «机器学习» (машинное обучение).

Инструменты и методы анализа данных

1. Инструменты и библиотеки: библиотеки обработки естественного языка в Python: NLTK, spacy, jieba
2. В изуализация данных: Math, Matplotlib

Результаты 1

Топ-50 высокочастотных слов во всей книге

Анализ английской версии: В английской версии текста часто встречаются слова "say" (более 1400 раз), "mr" (более 1200 раз) и "olive" (более 800 раз) [Рисунок 1]. Эти лексемы преимущественно сосредоточены на диалогах и характеристике персонажей. Частое употребление слова "say" указывает на важную роль диалогов в тексте; "mr" как форма обращения подчеркивает роль персонажей в разговорах; "olive" же обозначает конкретного персонажа, что дополнительно акцентирует внимание на информации о персонажах в тексте.

Анализ русской версии: В русской версии текста часто встречаются слова "сказать" (1068 раз), "мистер" (1057 раз) и "оливер" (872 раза) [Рисунок 1]. Эти лексемы аналогичны высокочастотным словам в английской версии и также сосредоточены на диалогах и обращениях к персонажам. Следует отметить, что в русской версии текста количество слов, появляющихся более 400 раз, на 7 меньше по сравнению с английской версией. Это явление может быть связано с

особенностями русского языка и стилем изложения текста, что отражает различия в частотном распределении лексики между языками.

Анализ китайской версии: В разных режимах сегментации высокочастотные лексемы китайского текста проявляют различные характеристики [рис.1]:

С полным сохранением словарного запаса: В этом анализе высокочастотные лексемы включают “说”("сказать") (1100 раз), “奥立弗”("Оливер")(959 раз), “一个”("один") (900 раз) и “不”("Нет") (776 раз). Эти слова отражают описание диалогов и персонажей. Особенно слова “说”("сказать") и “奥立弗”("Оливер") подчеркивают важность описания диалогов и персонажей в китайском тексте.

Только одиночные символы: В этом режиме высокочастотные лексемы включают “的”("из") (9505 раз), “一”("один") (6766 раз), “了”("Понятно") (4933 раз) и “他”("он") (4220 раз). Эти функциональные слова и местоимения встречаются в китайских текстах очень часто, демонстрируя общие элементы грамматической структуры китайского языка.

Только многосимвольные лексемы: В этом анализе высокочастотные лексемы включают “奥立弗”("Оливер") (959 раз), “一个”("один") (900 раз), “说道”("сказать") (563 раз) и “布尔”("логическое значение") (403 раз). Эти слова обычно связаны с конкретными персонажами и существительными, что указывает на использование большего количества существительных и глаголов при выражении конкретного содержания в китайском тексте.

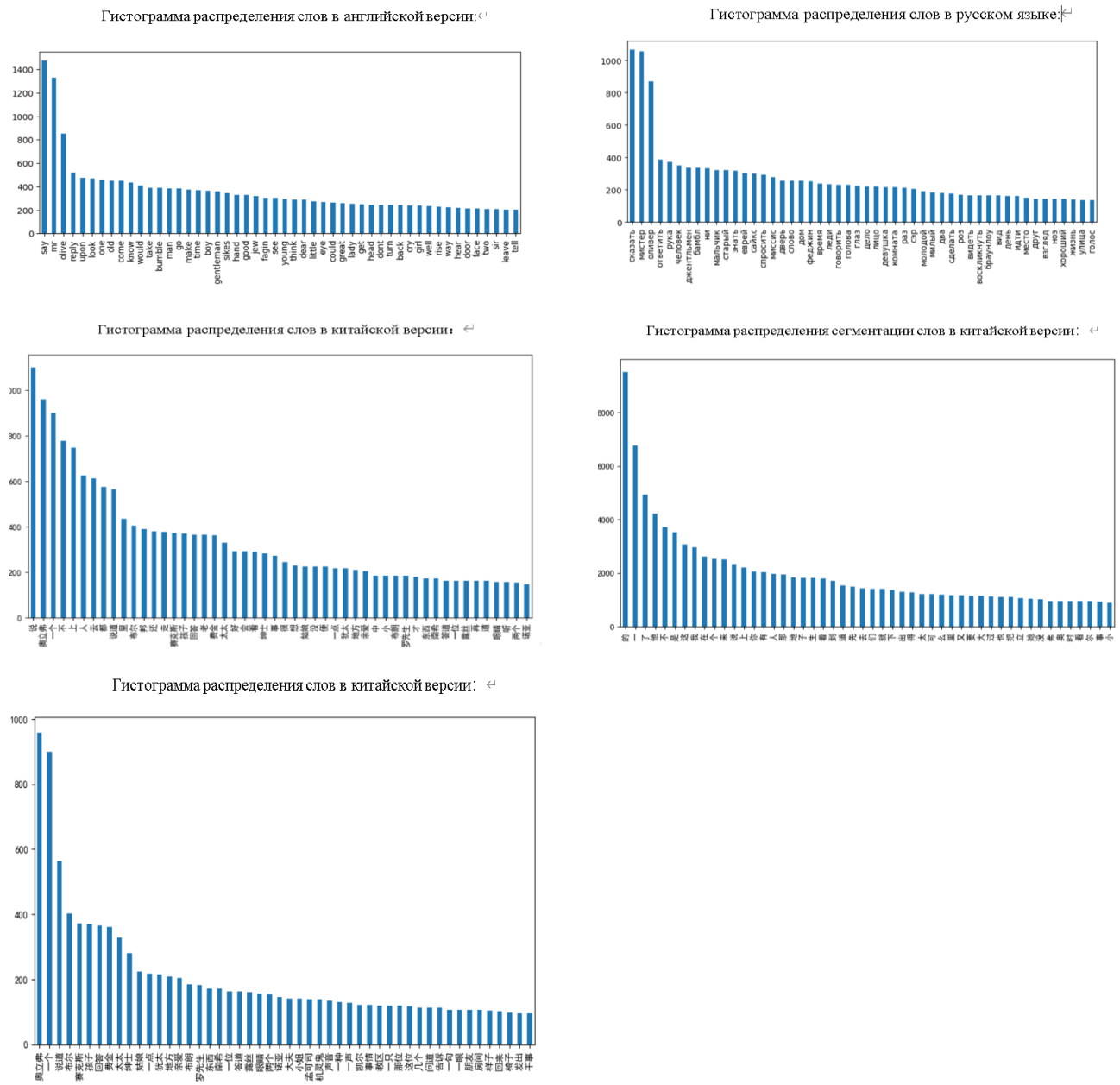


Рис. 1. Гистограмма распределения слов

Сравнительный анализ энтропии информации для книги

Анализ информационной энтропии, рассчитанной с помощью функций Хартли и Шеннона для текста на английском, русском и китайском языках, демонстрирует разнообразие языковых особенностей и их влияние на прогнозируемость информации [Таблица 1].

Для английского текста значения энтропии составляют H_Hartley = 9.0050 и H_Shanon = 7.4576. Эти показатели свидетельствуют о достаточно высокой предсказуемости текста на английском языке, что соответствует характерным особенностям языка, который имеет относительно фиксированный набор символов и грамматических структур.

В русской версии текста значения энтропии H_Hartley = 9.6366 и H_Shanon = 8.2633 немного выше, чем у английского текста. Это можно объяснить сложностью русской морфологии и большим количеством словоформ, что увеличивает вариативность информации и снижает предсказуемость текста.

Китайская версия текста демонстрирует наибольшие значения энтропии при учете полного словарного запаса (H_Hartley = 9.7566, H_Shanon = 8.3181) и многозначных символов (H_Hartley = 9.6856, H_Shanon = 8.5099). Эти результаты подчеркивают высокую сложность китайского языка и его богатый словарный запас, что делает текст менее предсказуемым и более вариативным по сравнению с английским и русским языками. В то же время, при анализе только отдельных символов, китайская версия показывает более низкие значения энтропии (H_Hartley = 8.0839, H_Shanon = 6.3134), что указывает на более высокую предсказуемость символов в данном контексте.

1. Таблица 1 - Энтропия информации к тексту книги

2.	3. H_Hartley	4. H_Shanon
5. Английская версия	6. 9.0050	7. 7.4576
8. Русская версия	9. 9.6366	10. 8.2633
11. Китайская версия(Полный словарный запас)	12. 9.7566	13. 8.3181

14. Китайская версия(Только отдельные символы)	15. 8.0839	16. 6.3134
17. Китайская версия(Только многозначные символы)	18. 9.6856	19. 8.5099

Проанализируйте распределение словарного запаса по разным главам

Сравнительный анализ распределения словарных единиц по главам в первых десяти главах различных языковых версий "Приключения Оливера Твиста" [рис. 2].

Горизонтальное сравнение (по главам):

Первая глава: Английская версия: На диаграмме видно, что распределение частоты использования слов достаточно разрозненное. Присутствуют слова, описывающие окружающую среду, такие как "gloomy" (мрачный) и "dreary" (угнетающий), а также вводящие персонажей, например, "boy" (мальчик) и "orphan" (сирота). Эти слова распределены относительно равномерно.

Русская версия: На диаграмме видны высокие столбцы для определенных слов, что указывает на более частое их использование в первой главе. Например, слова, описывающие атмосферу, такие как "темный" и "несчастный", занимают большую долю.

Китайская версия: Диаграмма показывает сбалансированное распределение слов, таких как "贫困" (бедность) и "孤儿" (сирота). Различия в долях слов незначительны.

Вторая глава: Английская версия: Появляются новые слова, связанные с развитием отношений, такие как "friend" (друг) и "enemy" (враг), или действия, продвигающие сюжет, например, "run" (бежать) и "hide" (прятаться). Их доля может изменяться в зависимости от сюжета. Русская версия: Стиль сохраняется, появляются новые слова, такие как "радость" и "горе", хотя изменения в доле слов относительно стабильны. Китайская версия: Распределение слов не меняется значительно, хотя могут появляться новые слова, такие как "惊讶" (удивление) и "疑惑" (сомнение), но их доля мала.

Третья глава и последующие: Английская версия: Изменения в распределении слов становятся более заметными. Появляются слова, отражающие сценарии, такие как "danger" (опасность) и "brave" (храбрый), или внутренний мир персонажей, например, "thoughtful" (задумчивый) и "worried" (обеспокоенный). Русская версия: Сохранение концентрации слов продолжается, могут появляться новые слова, например, "нервозность" (нервозность) в напряженные моменты, но изменения остаются плавными. Китайская версия: В распределении слов будут корректировки, появляются слова, описывающие эмоции, такие как "矛盾" (противоречие) и "和解" (примирение), с акцентом на многозначности и контексте, что поддерживает стабильность в различных главах.

Вертикальное сравнение (между языковыми версиями)

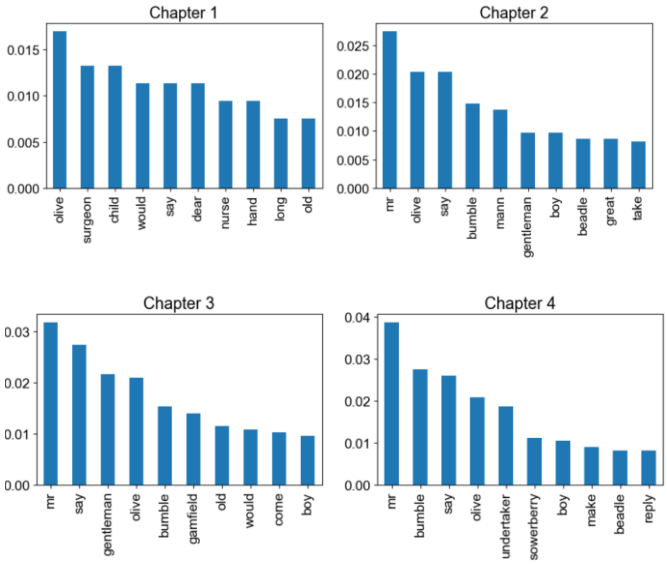
Богатство словарного запаса: При сравнении трех диаграмм можно заметить, что высота столбцов в английской версии значительно варьируется,

что указывает на более высокое богатство словарного запаса. Английская версия использует большое количество различных слов для выражения концепций и эмоций, что позволяет отображать разнообразие в распределении слов по главам. В русской версии высота столбцов изменяется менее резко, что указывает на менее высокое богатство словарного запаса. Здесь чаще повторяются определенные слова для выражения специфических эмоций и настроений, что делает распределение слов более концентрированным. В китайской версии количество столбцов и различия в высоте занимают промежуточное положение между английской и русской версиями. Китайский язык как бы сочетает богатство словаря с выбором некоторых специфических слов для усиления выражения, что делает распределение слов более сбалансированным.

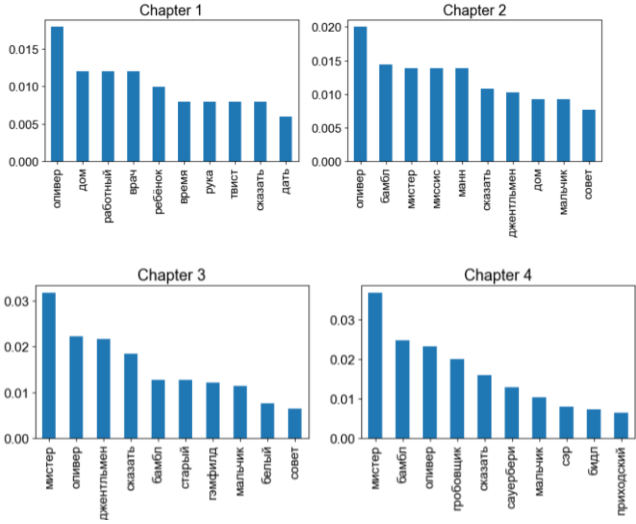
Степень отражения сюжета: Английская версия: Из-за разнообразия словарного запаса, изменения в распределении слов на диаграмме более значительны, что позволяет более детально отражать развитие сюжета и изменения в нем. Распределение слов в разных главах может точно отражать сюжетные повороты и эмоциональные изменения персонажей, например, в напряженных сценах связанные слова будут явно выделяться. Русская версия: Хотя распределение слов более сконцентрированное, диаграмма все же может отображать развитие сюжета. Например, в ключевых моментах сюжета могут появляться слова, частота использования которых увеличивается, подчеркивая основную тему и эмоции истории. Китайская версия: Диаграмма китайской версии показывает, что изменения в распределении слов более плавные, но различные комбинации слов и контекстуальное использование могут отражать

развитие сюжета. Например, при описании внутреннего мира персонажей используются слова с культурными коннотациями для передачи эмоций.

Аглийская версия



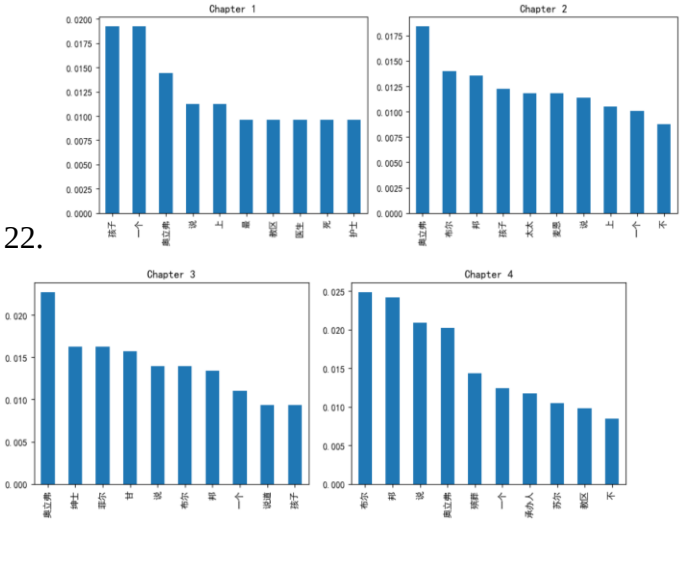
Русская версия



20. Китайская версия

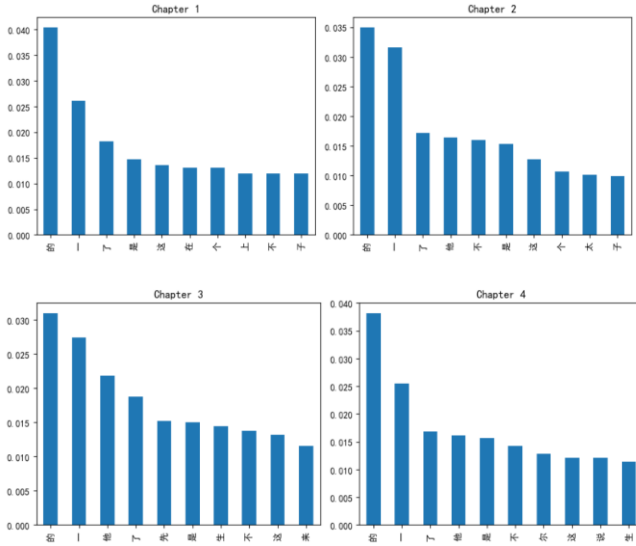
21. (Полный словарный запас)

22.



23. Китайская версия

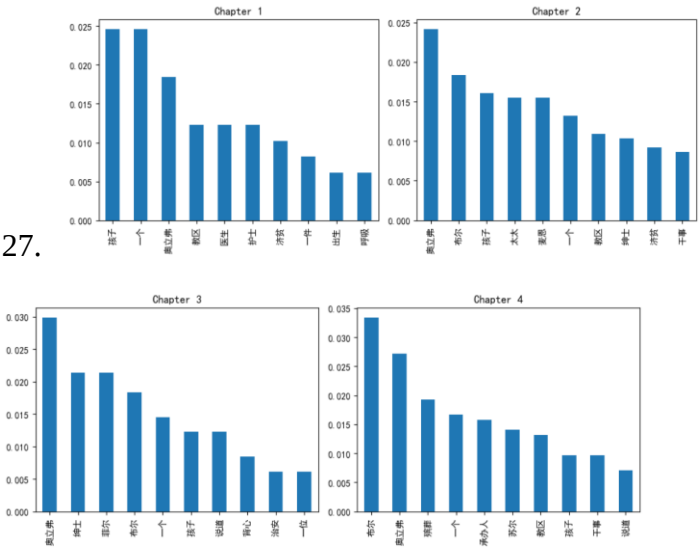
24. (Только отдельные символы)



25. Китайская версия

26. (Только многозначные символы)

28.

27. 

29. **Рис. 2.**Частотные диаграммы из 53 глав

Сопоставление изменений энтропии информации по главам

В результате вычислений энтропии информации по главам на трех языковых версиях и построения графиков на основе формул Хартли и Шеннона, были сделаны следующие наблюдения [рис.3]:

Общий анализ

1. Независимо от версии, тенденции сложности глав по статистике энтропии информации совпадают. Главы 26 и 38 имеют наивысшие значения энтропии, что указывает на максимальную сложность текста, вероятно, из-за увеличения плотности информации. Напротив, главы 36 и 45 показывают минимальные значения энтропии, что свидетельствует о меньшей сложности.

2. Изменения в сложности отражают стратегии изменения темпа повествования. Графики Хартли и Шеннона демонстрируют похожие тенденции, показывая значительные различия в энтропии между главами. При увеличении

номера главы нет четкой тенденции к росту или снижению энтропии, что говорит о нелинейности изменений сложности.

3. Методы расчета энтропии по Хартли и Шеннону показывают схожие тенденции, что подтверждает их согласование в отражении сложности текста по главам.

Анализ графиков для различных версий:

Английская версия: На графике видно, что значения энтропии информации в английской версии колеблются в определенном диапазоне. Это указывает на относительно стабильный уровень неопределенности информации в различных главах английского текста. Например, энтропия может колебаться в пределах от 6,2 до 6,8, что свидетельствует о стандартизированной грамматике и словаре. Эта стабильность может быть связана с тем, что английский язык как международный имеет относительно единообразную грамматику и словарный запас.

Русская версия: Энтропия информации в русской версии колеблется более значительно, значения распределены более разбросано. Это указывает на резкие изменения уровня неопределенности информации в разных главах русского текста. В некоторых главах может наблюдаться высокая энтропия, что отражает высокую сложность и неопределенность информации, в то время как в других главах значения энтропии могут быть низкими, показывая относительную простоту и определенность. Это может быть связано с особенностями русского языка, таким как сложная грамматическая структура и богатый словарный запас.

Китайская версия:

Полный словарный режим: Энтропия информации относительно высока, что свидетельствует о богатстве информации в китайском языке. В этом режиме каждый словарный элемент китайского текста рассматривается как единица информации, что приводит к высокой энтропии. Это отражает богатство и сложность китайского языка, так как его слова имеют многозначность и могут передавать сложные значения и эмоции. Графики Хартли и Шеннона могут демонстрировать некоторое сходство, но также могут различаться из-за различий в методах вычисления. Это показывает, что оба метода могут давать разные, но взаимодополняющие представления о сложности информации в китайском тексте.

Режим сегментации по одиночным символам: Энтропия информации заметно ниже, чем в полном словарном режиме. В этом режиме китайский текст разбивается на отдельные символы, каждый из которых имеет меньшую информационную ценность, что приводит к низкой энтропии. Это отражает особенности китайского языка, где одиночные символы требуют комбинации с другими символами для формирования значимых слов и предложений. График изменяется плавно, что свидетельствует о низкой сложности текста и малых различиях между главами.

Режим многосимвольных слов: Энтропия информации находится между полным словарным режимом и режимом одиночных символов. В этом режиме текст сохраняет многосимвольные слова, исключая одиночные символы и некоторые незначительные строки. Это сохраняет определенную информацию, снижая влияние одиночных символов, что приводит к средней энтропии. График

также имеет промежуточную тенденцию, показывая, что этот режим отражает сложность китайского текста, избегая при этом крайностей полного словарного и одиночного символного режимов.

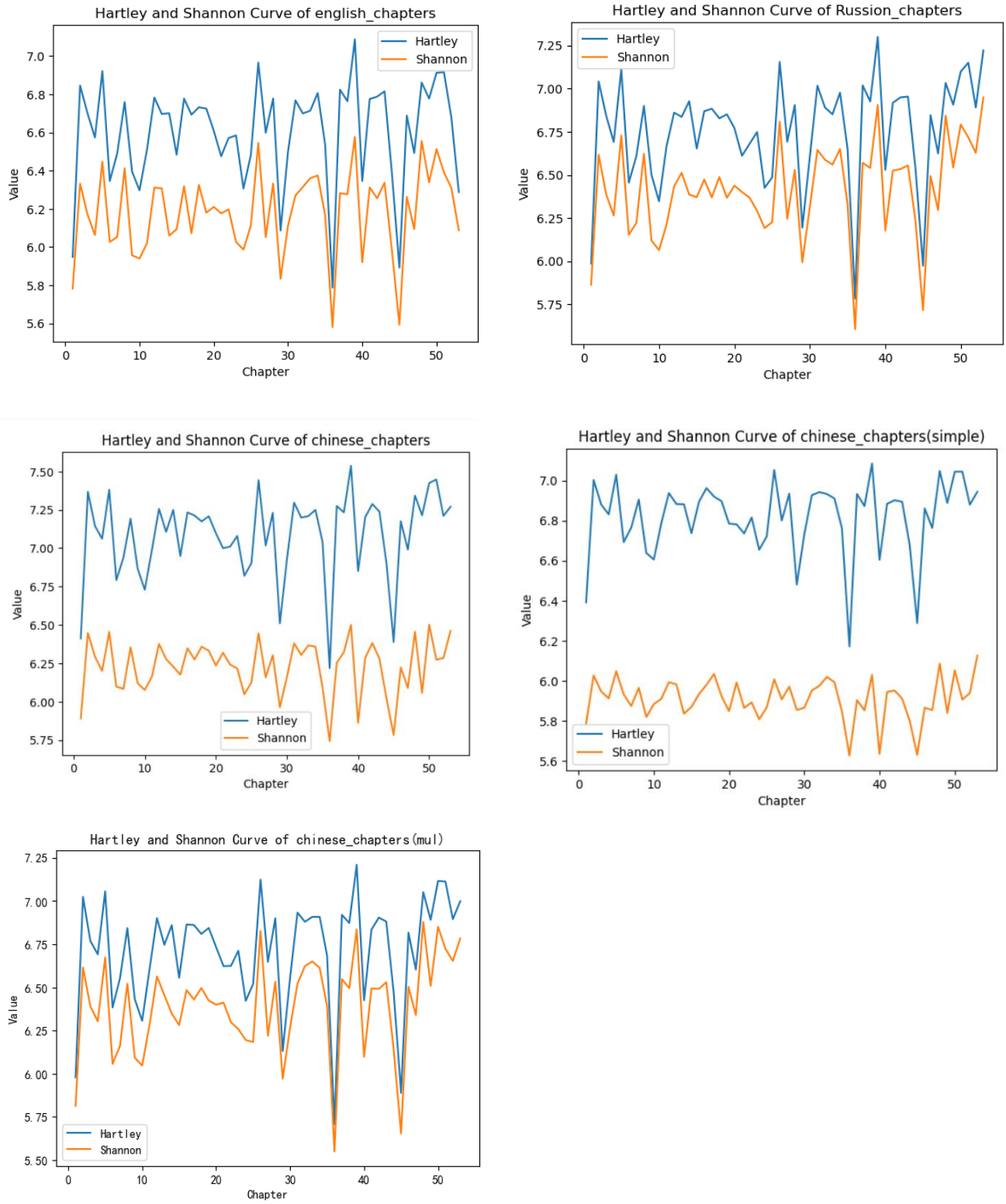


Рис. 3. График изменений информации (энтропии) по главам

Экспериментальное проектирование и сбор данных 2

Гипотеза

В контексте классификации китайских текстов выдвигается гипотеза, что при небольшом объеме данных сегментация на основе отдельных символов будет показывать лучшие результаты по сравнению с сегментацией на основе многосоставных фраз. Напротив, при большом объеме данных сегментация на основе многосоставных фраз, как ожидается, продемонстрирует более высокую эффективность по сравнению с сегментацией на основе символов.

Набор данных

Данное исследование использует набор данных для классификации текстов, состоящий из 40 различных романов, с целью создания основы для автоматической классификации глав романов. Выбранные романы охватывают как классические, так и современные произведения, что обеспечивает разнообразие и репрезентативность набора данных. Каждый роман разделен на главы, которые различаются по длине и структуре содержания, что позволяет модели изучать особенности различных повествовательных стилей и литературных форм. Главы помечены в соответствии с названиями соответствующих романов, что приводит к образованию 40 категорий.

Для обеспечения высокого качества и согласованности набора данных использовались различные методы предварительной обработки, включая удаление ненужных знаков препинания, стандартизацию текстовых форматов и

исключение шумовых данных. Количество текстовых примеров в каждой категории было тщательно спланировано, чтобы обеспечить достаточный размер выборки для обучения и оценки модели. В данном исследовании набор данных был разделен на обучающую и тестовую выборки, при этом 75% данных использовались для обучения модели, а 25% — для тестирования производительности модели. Обучающая выборка используется для обучения модели и настройки параметров, а тестовая — для оценки способности модели к обобщению на невиданных данных. В ходе эксперимента мы сравнивали точность, полноту, точность и другие показатели модели на тестовой выборке и анализировали изменения производительности модели в зависимости от различных вариантов выбора.

Извлечение функций

В этом исследовании мы использовали метод Term Frequency-Inverse Document Frequency (TF-IDF) для извлечения признаков. Мы сравнили два метода сегментации, как показано в Таблица 2:

Таблица 2 -МЕТОДЫ СЕГМЕНТАЦИИ

Метод сегментации	***Описание***
Сегментация на уровне символов	Рассматривает каждый символ как отдельный токен
Сегментация многосимвольных фраз	Рассматривает распространенные словосочетания или коллокации как единый токен.

Эти подходы позволяют нам оценить различия в текстовом представлении и их влияние на эффективность классификации.

Экспериментальная настройка

Эксперименты проводились с использованием Python, библиотеки jieba и библиотеки Scikit-learn. Процедуры обучения и тестирования следовали стандартным протоколам для обеспечения воспроизводимости. В качестве классификатора был выбран метод опорных векторов (SVM).

Результаты 2

Метод сегментации многословных фраз был применен к тестовому набору, и точность классификации достигла 88,0%, что значительно выше 55,0% для метода сегментации на уровне отдельных слов. Это свидетельствует о том, что метод сегментации многосимвольных слов более эффективен в захвате текстовых признаков и повышении точности классификации в данном проекте, как показано в Таблица 3.

При применении метода сегментации многосимвольных слов большинство категорий показали высокие значения точности, полноты и F1-меры, особенно для категорий 1, 4, 8, 10, 11, 26, 29, 30, 34, 36, 39 и 40, все из которых достигли идеального результата 1,00. Даже для категорий, которые продемонстрировали относительно слабые результаты, таких как 2, 3, 5, 12, 14, 16, 21, 22, 23, 24, 25, 31, 35, 37 и 38, показатели оставались на разумном уровне, без случаев, когда они были равны нулю.

В методе сегментации на уровне символов многие категории показали значения точности, полноты и F1-меры равные 0,00, что указывает на полную неспособность этого подхода корректно классифицировать эти категории. Например, категории 2, 7, 8, 10, 14, 16, 18, 19, 20, 22, 24, 26, 28, 29, 30, 31, 34, 36 и 40 продемонстрировали данную проблему. Для других категорий с ненулевыми метриками значения обычно были ниже, чем у соответствующих категорий в методе сегментации многосимвольных слов.

Таблица 3 -ЭКСПЕРИМЕНТАЛЬНЫЕ РЕЗУЛЬТАТЫ

Number	ЭКСПЕРИМЕНТАЛЬНЫЕ РЕЗУЛЬТАТЫ						
	Precision (multi)	***Recall (multi)***	***F1-Score (multi)***	***Precision (single)***	***Recall (single)***	***F1-Score (single)***	***Support***
1	0.00	0.91	0.95	0.91	0.91	0.91	11
2	1.00	1.00	1.00	1.00	1.00	1.00	21
3	1.00	0.50	0.67	0.00	0.00	0.00	4
4	1.00	0.61	0.76	1.00	0.39	0.56	18
5	1.00	0.78	0.88	1.00	0.78	0.88	9
6	1.00	0.96	0.98	1.00	0.96	0.98	28
7	0.94	0.97	0.98	0.00	0.00	0.00	18
8	0.00	0.00	0.00	0.00	0.00	0.00	1
9	0.00	0.00	0.00	0.43	1.00	0.60	36
10	1.00	0.97	0.99	0.00	0.00	0.00	8
11	1.00	1.00	1.00	1.00	0.25	0.40	8
12	0.67	0.80	0.80	1.00	0.33	0.50	3
13	0.00	0.00	0.00	0.30	0.36	0.51	11
14	0.00	0.00	0.00	0.00	0.00	0.00	3

15	0.97	0.98	0.98	0.16	0.28	0.20	30
16	0.58	1.00	0.74	0.00	0.00	0.00	7
17	0.00	0.00	0.00	0.92	0.73	0.81	15
18	0.83	0.91	0.91	0.00	0.00	0.00	18
19	1.00	0.80	0.89	0.00	0.00	0.00	5
20	1.00	0.86	0.92	0.00	0.00	0.00	7
21	0.84	0.80	0.82	0.84	0.80	0.82	20
22	0.00	0.00	0.00	0.00	0.00	0.00	5
23	1.00	0.64	0.78	0.69	0.79	0.73	14
24	1.00	0.75	0.86	0.00	0.00	0.00	4
25	0.31	0.96	0.46	0.64	0.30	0.41	23
26	1.00	1.00	1.00	0.00	0.00	0.00	15
27	0.96	0.96	0.96	0.90	1.00	0.95	28
28	1.00	0.92	0.96	0.00	0.00	0.00	13
29	1.00	1.00	1.00	0.00	0.00	0.00	1
30	0.83	0.91	0.91	0.00	0.00	0.00	12
31	0.90	0.94	0.97	0.00	0.00	0.00	10
32	0.94	0.97	0.97	0.89	0.94	0.91	17
33	1.00	1.00	1.00	1.00	1.00	1.00	11
34	1.00	1.00	1.00	0.00	0.00	0.00	7
35	0.95	0.69	0.80	0.79	0.88	0.84	26
36	1.00	1.00	1.00	0.00	0.00	0.00	3
37	1.00	1.00	1.00	1.00	0.23	0.38	38
38	1.00	1.00	1.00	1.00	0.33	0.50	15
39	1.00	1.00	1.00	1.00	0.25	0.40	4
40	1.00	1.00	1.00	0.00	0.00	0.00	3
accuray(mul)						0.88	512
accuray(sim)						0.55	512

Для исследования влияния двух методов сегментации на производительность классификации при различных объемах данных, мы проанализировали изменения точности сегментации многосимвольных слов (TF-IDF 1) и сегментации на уровне символов (TF-IDF 2) с разным количеством обучающих примеров, как показано на Рис. 4.

На рисунке видно, что по мере увеличения числа обучающих примеров с 75 до 1575 метод сегментации многосимвольных слов стабильно демонстрирует значительное преимущество. Точность метода TF-IDF 1 постоянно улучшается с увеличением объема данных, что свидетельствует о том, что многосимвольная сегментация позволяет лучше использовать большие наборы данных для обучения текстовых признаков, сохраняя больше семантической информации и контекстных связей. Это позволяет модели более точно понимать смысл текста и улучшать точность классификации. В отличие от этого, точность метода сегментации на уровне символов (TF-IDF 2) остается относительно низкой и демонстрирует менее выраженный рост по сравнению с методом многосимвольной сегментации. Сегментация на уровне символов слишком мелко разделяет текст, что может привести к утрате важной семантической информации, что затрудняет модели точное захватывание текстовых признаков, особенно по мере увеличения объема данных, когда ограничения данного метода становятся более очевидными.

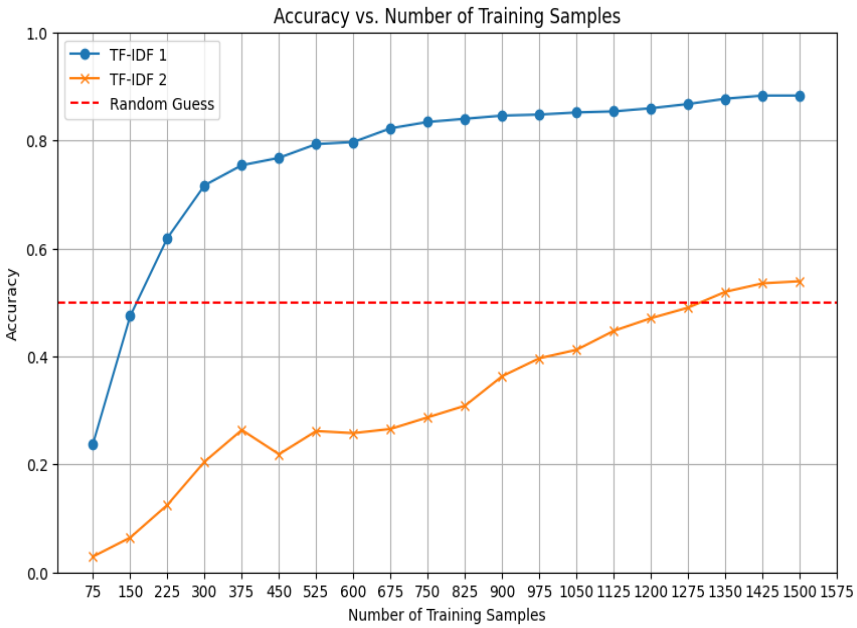


Рис. 4 Изменение точности сегментации многосимвольных слов (TF-IDF 1) и сегментации односимвольных слов (TF-IDF 2) при различном количестве обучающих выборок

Заключение

Это исследование использует методы частотного анализа, расчета энтропии и статистического анализа для изучения языковых характеристик и их влияния на перевод одного литературного произведения на английский, русский и китайский языки. Результаты показывают, что выбор лексики и частоты слов в переводах отличается, что отражает особенности стиля каждого языка. Английская версия имеет относительно стабильную энтропию, русская версия демонстрирует большие колебания энтропии, а китайская версия имеет высокую энтропию при использовании полной лексики и низкую при использовании отдельных символов. Анализ сложности текста выявил различия между главами произведения.

После проведения экспериментов по классификации текста с использованием различных методов сегментации, наше исследование ясно показало, что сегментация на уровне многословных единиц значительно превосходит сегментацию на уровне символов при классификации литературных произведений. Этот вывод подчеркивает важность использования более сложных методов сегментации для повышения точности классификации текста. Эффективно захватывая тонкости фраз и контекстные значения, сегментация на уровне многословных единиц улучшает общую производительность классификационных моделей. Эти результаты не только вносят вклад в существующие знания в области обработки естественного языка, но и открывают новые перспективы для будущих исследований и практического

применения в различных областях, способствуя более эффективной обработке китайских текстов.

Список источников

1. Dickens, Ch. (2000). Oliver Twist. Hertfordshire: Wordsworth Editions.

2. Gary Lupyan, Rick Dale, Why Are There Different Languages? The Role of Adaptation in Linguistic Diversity,Trends in Cognitive Sciences,Volume 20, Issue 9, 2016, Pages 649-660, ISSN 1364-6613, https://doi.org/10.1016/j.tics.2016.07.005.

3. Shaw, R. D. (1987). The Translation Context: Cultural Factors in Translation. Translation Review, 23(1), 25–29. https://doi.org/10.1080/07374836.1987.10523398

4. W. A. Qader, M. M. Ameen and B. I. Ahmed, "An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges," 2019 International Engineering Conference (IEC), Erbil, Iraq, 2019, pp. 200-204, doi: 10.1109/IEC47844.2019.8950616.

5. G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. Commun. ACM 18, 11 (Nov. 1975), 613–620. https://doi.org/10.1145/361219.361220

6. Dumais, S.T. (2004). Latent Semantic Analysis. Annual Review of Information Science and Technology (ARIST), 38, 189-230. Retrieved September 9, 2024 from https://www.learntechlib.org/p/97537/.

7. Qaiser, S.; Ali, R. Text mining: Use of TF-IDF to examine the relevance of words to documents. Int. J. Comput. Appl. 2018, 181, 25–29.

8. L. Ma and Y. Zhang, "Using Word2Vec to process big text data," 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, USA, 2015, pp. 2895-2897, doi: 10.1109/BigData.2015.7364114.

9. Danyal, M.M., Khan, S.S., Khan, M. et al. Proposing sentiment analysis model based on BERT and XLNet for movie reviews. Multimed Tools Appl 83, 64315–64339 (2024). https://doi.org/10.1007/s11042-024-18156-5

10. S. Black, S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, et al., "GPT-NeoX-20B: An open-source autoregressive language model", arXiv:2204.06745, 2022.

11. Aditi, S. Shandilya, N. Bansal and S. Mala, "An Evaluation of Word Frequency Techniques for Text Summarization Using Sentiment Analysis Approach," 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2020, pp. 397-403, doi: 10.1109/Confluence47617.2020.9058139.

12. Grefenstette, G. (1999). Tokenization. In: van Halteren, H. (eds) Syntactic Wordclass Tagging. Text, Speech and Language Technology, vol 9. Springer, Dordrecht. https://doi.org/10.1007/978-94-015-9273-4_9

13. Steven B, Ewan K, Edward L. (2009). NLTK Natural Language Processing with Python. California: O 'Reilly Media.

14. Ryan Spring, Matthew Johnson,The possibility of improving automated calculation of measures of lexical richness for EFL writing: A comparison of the LCA, NLTK and SpaCy tools,System,Volume 106,2022,102770,ISSN 0346-251X, https://doi.org/10.1016/j.system.2022.102770.

15. H. Liu, P. Gao and Y. Xiao, "New Words Discovery Method Based On Word Segmentation Result," 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), Singapore, 2018, pp. 645-648, doi: 10.1109/ICIS.2018.8466490.

16. R. Pramana, Debora, J. J. Subroto, A. A. S. Gunawan and Anderies, "Systematic Literature Review of Stemming and Lemmatization Performance for Sentence Similarity," 2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA), Yogyakarta, Indonesia, 2022, pp. 1-6, doi: 10.1109/ICITDA55840.2022.9971451.

17. Hartley, R.V.L. Transmission of Information. Bell Syst. Tech. J. 1928, 7, 535–563.

18. Shannon, C.E. A mathematical theory of communication. Bell Syst. Tech. J. 1948, 27, 379–423.

19. Dickens, Ch. (2010). Oliver Twist. Nanjing: Yilin Publishing House.

20. Charles Dickens (1998). Adventures The Adventures of Oliver Twist. Moscow: Publishing House "KhudLit".