Вычислительные технологии, 2024, том 29, № 2, с. 113–120. © ФИЦ ИВТ, 2024
Computational Technologies, 2024, vol. 29, no. 2, pp. 113–120. © FRC ICT, 2024
ISSN 1560-7534
eISSN 2313-691X

Федеральный исследовательский центр информационных и вычислительных технологий

Новосибирский государственный университет

ОБЪЕДИНЕННЫЙ СЕМИНАР

ИНФОРМАЦИОННО-ВЫЧИСЛИТЕЛЬНЫЕ ТЕХНОЛОГИИ В ЗАДАЧАХ ФИЛОЛОГИИ И КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

Руководители: д-р техн. наук, канд. филол. наук О.Ю. Кожемякина, д-р техн. наук, доцент В. Б. Барахнин

Аннотации докладов за 2023 г.

Информационная энтропия поэтических текстов в задачах стилеметрии

О. Ю. Кожемякина
Федеральный исследовательский центр информационных и вычислительных технологий, Новосибирск (21.02.2023)

Одним из методов количественного анализа текста является его представление в виде временно́го ряда с последующим изучением информационной энтропии текста. Исследование авторского стиля на основании энтропийных характеристик — перспективное направление в области анализа информации поэтического текста. В рамках работы: проведены расчеты значений энтропии первого, второго и третьего порядка для корпусов стихотворений А.С. Пушкина и других поэтов пушкинской эпохи; получены математическое ожидание и дисперсия подсчитываемого в каждой серии вида энтропии; начаты расчеты, имеющие целью выяснить значимость описанных выше энтропийных характеристик для определения авторского стиля; реализовано программное приложение, автоматически извлекающее статистическую информацию, потенциально применимую в задачах выявления особенностей авторского стиля, из русскоязычных поэтических текстов и их транскрипций; извлечены статистические данные из стихотворений А.С. Пушкина и других авторов, которые могут стать основой стилеметрической классификации авторов по энтропийным признакам.

Автоматизированный анализ метроритмических характеристик поэтических текстов на русском языке

И. В. Кузнецова
Федеральный исследовательский центр информационных и вычислительных технологий, Новосибирск (28.02.2023)

В докладе представлено диссертационное исследование на тему автоматизированного анализа метроритмических характеристик поэтических текстов на русском языке.

Описаны алгоритмы и реализация программного модуля, отвечающего за структурный уровень анализа поэтического текста:

- алгоритм определения метра и стопности;
- алгоритм определения типа рифмовок в строфе стихотворения.

Представлены математическая модель фактуры поэтического текста для формальной модели фактуры, разработанной О.Ю. Кожемякиной, и ее программная реализация. Алгоритмы реализованы на языке Python и протестированы на корпусах стихотворений А.С. Пушкина и А.К. Толстого.

Сравнительный анализ значимости синтаксических признаков текста при оценке его сложности

И. А. Смаль
Новосибирский государственный университет, Новосибирск (14.03.2023)

Представлены результаты исследования возможной точности предсказания сложности текста на русском языке методами градиентного бустинга и случайного леса с использованием синтаксических признаков. Работа проводилась на двух корпусах текстов — художественной и учебной литературы. Анализ показывает, что использования только синтаксических характеристик текста недостаточно для выявления сложности его понимания. Для текстов разных стилей наиболее значимые признаки различны: для художественной литературы самый значимый признак — среднее количество зависимостей “conjunction”, для учебной литературы — среднее количество зависимостей “nominal modifier” и “adjective modifier”. В дальнейшем предполагается построение более сложных моделей с использованием также лексических признаков.

Использование подхода переноса обучения языковых моделей для извлечения именованных сущностей и отношений между ними

Н. А. Шварц, Т. В. Батура
Новосибирский государственный университет, Новосибирск (21.04.2022)

Приведены результаты исследования подхода переноса обучения языковых моделей для извлечения именованных сущностей и отношений между ними. В частности, был рассмотрен и проанализирован сквозной (end-to-end) метод решения поставленной задачи. Метод предполагает, что входными данными являются неподготовленные текстовые данные, а в качестве результата необходимо получить границы именованных сущностей, а также тип (класс) отношения между ними. Таким образом, в рамках одной постановки решаются две крупнейшие задачи области обработки естественных языков: named entity recognition (NER) и relation classification (RE). В дополнение к данной формулировке были рассмотрены условия обучения языковых моделей без примеров (zero-shot learning, ZSL). Такая постановка позволяет рассмотреть более сложные применения исследованного решения, в частности осуществление переноса знаний с одной предметной области на другую. В работе представлены анализ новой постановки задачи, не рассмотренной ранее, предлагаемое решение и анализ проведенных экспериментов. Показано, что предлагаемое решение перспективно в случае сложной комбинированной задачи, однако требует дальнейших исследований.

Разработка и реализация программного приложения для сравнительного анализа статистических характеристик поэтических текстов

Э. Д. Кожемякина
Федеральный исследовательский центр информационных и вычислительных технологий, Новосибирск (28.03.2023)

В ФИЦ ИВТ разработано программное приложение для статистического исследования фонетической транскрипции поэтических текстов. Поставлена задача расширения функционала приложения: интеграция с базой данных, используемой в системе комплексного анализа поэтических текстов, а также построение сводных графиков для сравнительного анализа фонетических характеристик как конкретных поэтических текстов, так и их корпусов. В докладе представлено реализованное программное приложение, удовлетворяющее сформулированным выше требованиям и предназначенное для внедрения в программную систему комплексного анализа поэтических текстов.

Использование современных сетей Хопфилда для обучения нейросетевых алгоритмов распознавания речи на неразмеченных данных

Д. В. Гребенкин
Новосибирский государственный университет, Новосибирск (04.04.2023)

В работе представлена гипотеза: использование “плотной” ассоциативной памяти вместо механизма внимания, которая позволит сделать интегральную модель распознавания речи устойчивой к шумам, реверберационным эффектам и особенностям речи (диалектам, нарушениям произношения). Для проверки данной гипотезы создано две модели на основе нейросетевого алгоритма Wav2Vec2: с механизмом внимания (w2v2-classic-tiny) и с современными сетями Хопфилда (w2v2-hopfield-tiny). Результаты тестирования показали, что обученная на небольшом количестве данных модель со слоями “плотной” ассоциативной памяти имеет лучшее качество, чем модель со стандартной архитектурой. Полученные результаты интерпретированы на основе анализа различных звуков, предсказанных моделью, с использованием UMAP-проекций и автоматического выравнивания.

Использование нейронных сетей для определения тематических характеристик русских поэтических текстов на основании лексических признаков

Н. С. Каминский
Новосибирский государственный университет, Новосибирск (11.04.2023)

Важной и весьма сложной задачей автоматизации комплексного анализа поэтических текстов является разработка алгоритмов определения их тематических характеристик. Перспективным представляется подход, связанный с использованием методов машинного обучения, прежде всего нейросетей. В рамках представленной работы уточнен состав корпуса стихотворений известных русских поэтов о временах года, который выступал в роли обучающей и тестовой выборки в сравнительных экспериментах. Для классификации пейзажной лирики по признаку “время года” использованы различные архитектуры нейронных сетей, в том числе Word2Vec и BERT. Наилучшие результаты

получены с применением BERT. В дальнейшем предполагается провести серию экспериментов с использованием fastText.

Проблемы конвертации данных в форматах JSON и XML

П. А. Ловкий
Новосибирский государственный университет, Новосибирск (25.04.2023)

В докладе представлен обзор основных проблем, связанных с преобразованием данных. Для интеграции между web-системами, использующими различный формат данных в качестве основного (JSON или XML), необходим механизм конвертации этих форматов друг в друга. Существующие преобразователи JSON-данных в XML-данные не предоставляют интерфейса для проверки выходных данных на соответствие определенной структуре, соответствие определенных значений конкретным типам и т. д., что может привести к потере или некорректности некоторых данных после редактирования. В частности, такая проблема возникает при "круговом" преобразовании. Кроме того, необходим механизм поддержки пространств имен в XML для задания различных ограничений. В докладе рассмотрены существующие популярные решения, их преимущества и недостатки, а также приведены результаты некоторых исследований по данной теме.

Методы машинного обучения в анализе интернет-комментариев, содержащих оценку высших учебных заведений

Г. Д. Перетолчин
Новосибирский государственный университет, Новосибирск (02.05.2023)

Рассмотрены методы представления комментариев в сети, алгоритмы их предобработки и модели машинного обучения, используемые для классификации текстов. Основной задачей является классификация комментариев по тональности, источник анализируемого материала — интернет-ресурс `https://tabiturient.ru`.

Разработана методика анализа интернет-комментариев, содержащих оценку высших учебных заведений. Оценка точности методики позволяет сделать вывод: наивный байесовский классификатор дает достаточно хорошие результаты, однако лучшие результаты получены с использованием LSTM-модели.

Применение современных сетей Хопфилда для улучшения нейронных языковых моделей

А. К. Берзин
Новосибирский государственный университет, Новосибирск (16.05.2023)

Нейронные языковые модели успешно применяются в задачах обработки естественного языка, однако с увеличением числа параметров улучшение качества работы модели в зависимости от масштабирования значительно замедляется. В докладе выдвигается гипотеза: сети Хопфилда могут лучше моделировать ассоциативную память, чем нейронные сети прямого распространения. Для проверки предположения проводится сравнение двух BERT-подобных языковых моделей: исходной и модифицированной сети Хопфилда. Предлагается ряд оптимизаций, которые могут быть применены к другим нейронным языковым моделям.

Организация хранения корпусов поэтических текстов в информационных аналитических системах с учетом специфики предметной области

О. Ю. Кожемякина, Н. А. Шашок, Э. Д. Кожемякина
Федеральный исследовательский центр информационных и вычислительных технологий, Новосибирск (10.10.2023)

Информационная система, как соответствующий компонент программной системы, объединяет разнородную информацию о результатах анализа поэтических текстов, структура которых иерархична согласно их языковой природе. Вопрос иерархии текста равнозначно важен для процесса его анализа и для хранения корпусов текстов, что необходимо учитывать при разработке информационных систем, предназначенных для хранения и обработки текстов на естественном языке. Хранилище текстов является, как правило, центральным компонентом информационных аналитических систем и либо проектируется как база данных, либо представляет собой неструктурированный набор данных. В процессе концептуального проектирования хранилища корпусов поэтических текстов, с учетом специфики объектов предметной области, обосновано целесообразное использование двух систем хранения и поиска данных: реляционной базы данных для хранения связей между объектами в системе, а также объектов, не являющихся частью корпуса, и хранилища файлов с инструментом полнотекстового поиска в корпусе текстов, что повышает качество анализа текстов и расширяет возможности применения системы в целом.

Модуль анализа метроритмических и строфических характеристик русских поэтических текстов

И. В. Кузнецова
Федеральный исследовательский центр информационных и вычислительных технологий, Новосибирск (17.10.2023)

В докладе представлен модуль комплексной системы автоматизированного анализа русских поэтических текстов, производящий анализ метроритмических и строфических характеристик стиха. Построена математическая модель фактуры русского стиха, основанная на модифицированном алгоритме И.А. Пильщикова и разработанной модели рифмы. На основе этой математической модели разработаны и реализованы алгоритмы определения метроритмических и строфических характеристик русских поэтических текстов, а также алгоритм определения фактуры, на основе которого реализована программа на языке Python, определяющая фактуру входного стихотворения.

Разработка структуры документов с пересекающейся сегментацией в системе Elasticsearch

Н. А. Шашок, Э. Д. Кожемякина
Федеральный исследовательский центр информационных и вычислительных технологий, Новосибирск (17.10.2023)

В процессе автоматического создания словарей авторского языка и конкордансов возникает задача определения контекста употребления лексики, при этом строфы и строки явно связаны иерархическими отношениями, однако строки и предложения, а также предложения и строфы иерархических отношений не имеют. Форматы хранения

и передачи текстовых данных имеют, как правило, иерархичный характер, таким образом, практический интерес представляет разработка принципов структуризации текстов с учетом выявленных пересекающихся сегментов в рамках задачи поиска контекста с предварительно заданным уровнем сегментации. В докладе представлена структура документов JSON, использование которой в рамках индекса Elasticsearch позволяет осуществлять поиск контекста употребления лексики в корпусе поэтических текстов, хранящегося в индексе.

Разработка методики выявления ложных отзывов на сайтах торговых сетей

В. А. Лисин
Новосибирский государственный университет, Новосибирск (31.10.2023)

В докладе рассматривается актуальная проблема распознавания ложных (сгенерированных компьютером или одним и тем же пользователем под разными именами) отзывов, вводящих в заблуждение потребителей и компании, предоставляющие товары и услуги. Сложность задачи заключается в разнородности исходных данных, требующих глубокого анализа для установления ложности отзыва. Лежащий в основе массива данных корпус необработанных отзывов, собранных с различных платформ (маркетплейсов и агрегаторов), не имеет и не может иметь готовой разметки, поэтому необходимо максимально упростить процесс разметки отзывов. Предлагается разработать комплексный метод анализа отзывов, приведена возможная схема решения данной задачи с использованием методов машинного обучения и определены планы по оптимизации предложенного подхода.

Автоматическая разметка документов для научно-просветительского ресурса “Пушкин Цифровой”

Н. Н. Тесля
Санкт-Петербургский Федеральный исследовательский центр РАН, Санкт-Петербург (14.11.2023)

Создание научно-просветительских ресурсов для популяризации и широкого распространения авторского наследия требует автоматизации процессов, связанных с наполнением ресурса материалами и метаданными по материалам. Одним из таких процессов является разметка текста для загруженных документов. В процессе разметки осуществляется поиск значимых сущностей в тексте, с помощью которых производится классификация, расширяется количество возможных связей документов и упрощается поиск документов в научно-просветительском ресурсе. Заполнение метаданных возможно на основе автоматического поиска их по базе знаний с использованием названия и типа документа. Для решения представленных задач предлагается метод на основе использования мультиязычной модели BERT, дополнительно обученной на корпусе размеченных текстов из Пушкинской энциклопедии, а также использование семантических запросов к порталу Викиданные.

Информационная энтропия поэтического текста: история, эксперименты, перспективы

О. Ю. Кожемякина, Н. А. Шашок
Федеральный исследовательский центр информационных и вычислительных технологий, Новосибирск (28.11.2023)

Информационная составляющая поэтического текста является нетривиальным объектом для исследования, поскольку поэтический текст — это сложная многомерная структура. Количественные методы анализа текстовой информации имеют давнюю традицию применения в отечественной науке; в современных исследованиях очевиден интерес к тематике, связанной с вычислением энтропийных характеристик и последующим применением полученных данных в задачах анализа текстовой информации на естественном языке. Результаты вычислений буквенной, звуковой и "эмоциональной" энтропии в рамках прикладной задачи исследования энтропийных характеристик и продолжение экспериментальных расчетов с использованием расширенной тестовой выборки позволяют говорить об обоснованности применения энтропийных критериев в задачах определения авторского стиля, при этом перспективным направлением становится изучение фонетической составляющей информативности поэтического текста, а также учет статистических данных по знакам препинания. Современные информационные технологии с опорой на методы классической математики и информатики позволяют выявить скрытые отношения и связи в статическом пространстве поэтического текста, а также неявные процессы в динамическом плане восприятия поэтического произведения. В широком аспекте информационная энтропия поэтического текста становится одним из основных параметров для определения информативности поэтического текста и, соответственно, вспомогательным инструментом понимания культурного кода автора и читателя.

Применение программного комплекса "Гипертекстовый поиск слов-спутников в авторских текстах" в филологических исследованиях: итоги и перспективы

Л. В. Павлова, И. В. Романова
Смоленский государственный университет, Смоленск (05.12.2023)

В докладе представлены принципы работы оригинального программного комплекса "Гипертекстовый поиск слов-спутников в авторских текстах", предназначенного для автоматического выделения лексических комбинаций в поэтических текстах, а также два направления изучения лексических комбинаций — "корпусное" и "тематическое". Под лексическими комбинациями понимаются наборы лексем, повторяющиеся в разных текстах одного автора или в стихотворениях разных авторов, при этом между лексемами в составе комбинации может не быть никаких грамматических или стиховых связей. Предтекстовая природа этих бесструктурных образований не препятствует их роли на этапе формирования текста, где лексические комбинации устанавливают грамматические, фонетические, ритмико-синтаксические отношения и обретают семантическую специфику, уникальную для каждого автора. Появление повторяющихся лексических комбинаций — дифференциальная особенность поэтической речи, поэтому они могут быть еще одним способом описания пространства поэтического текста.

Методы глубокого обучения в расшифровке тибетских старопечатных книг

О. С. Ринчинов
Институт монголоведения, буддологии и тибетологии СО РАН, Улан-Удэ (12.12.2023)

Тибетская литература является одной из наиболее древних и развитых литературных традиций, уходящих корнями в VII в. Она до сих пор определяет культурную и религиозную жизнь на огромных пространствах Внутренней Азии, включая восточные регионы России. В докладе рассказывается о современных подходах к цифровизации тибетоязычного письменного наследия, связанных с применением методов искусственного интеллекта, и первых полученных в этой области результатах.

Место и время проведения заседаний: по вторникам, в 17:30,
конференц-зал Федерального исследовательского центра информационных и вычислительных технологий
Адрес: просп. акад. Лаврентьева, 6, Новосибирск, 630090
Секретарь семинара: аспирант ФИЦ ИВТ Наталья Александровна Шашок
e-mail: n.shashok@alumni.nsu.ru
Интерактивная заявка доклада:
http://www.ict.nsc.ru/ru/education/seminar/seminar-page-lingv