Вычислительные технологии, 2023, том 28, № 1, с. 118–124. © ФИЦ ИВТ, 2023
Computational Technologies, 2023, vol. 28, no. 1, pp. 118–124. © FRC ICT, 2023
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, доцент В. Б. Барахнин, канд. филол. наук О. Ю. Кожемякина

Аннотации докладов за 2022 г.

Проект РНФ “Разработка и реализация информационной системы многоуровневого исследования стихотворных текстов” (2019–2021): задачи и результаты

О. Ю. Кожемякина
Федеральный исследовательский центр информационных и вычислительных технологий, Новосибирск (15.03.2022)

В ходе выполнения проекта РНФ были разработаны и реализованы алгоритмы, предназначенные для автоматизированного извлечения из поэтических текстов их характеристик, относящихся к различным структурным уровням (метрика, ритмика, фонетика, лексика, грамматика, литературный стиль, тематика, литературный жанр), и исследования их взаимозависимости. На основе разработанных алгоритмов создана программная система комплексного анализа русских поэтических текстов, позволяющая освободить исследователей-стиховедов от рутинной работы, расширить круг анализируемых произведений и получить новые научные результаты, связанные с исследованием взаимозависимостей структур различных уровней русских поэтических текстов.

Программная система комплексного анализа русских поэтических текстов находится в открытом доступе в сети Интернет по адресу `http://poem.ict.nsc.ru/system`.

Формальный анализ аргументационных структур в научных и научно-популярных текстах

И. С. Пименов
Новосибирский государственный университет, Новосибирск (22.03.2022, 29.03.2022)

Доклад посвящен автоматическому анализу аргументации в научных и научно-популярных текстах на русском языке.

Рассматриваются задачи и методы: 1) аргументационной разметки текстов; 2) извлечения аргументативных утверждений; 3) выявления приемов аргументации, повторяющихся в разных текстах; 4) характеризации текстов по способу убеждения (выраженности пафоса и этоса).

Предложенные методы апробируются на коллекции научно-популярных и научных (по компьютерным технологиям и лингвистике) статей.

На основе экспериментальных результатов описывается специфика построения доказательств в научных и научно-популярных текстах.

Разработка и реализация программного приложения для статистического исследования транскрипции поэтических текстов

Э. Д. Кожемякина

Муниципальное бюджетное общеобразовательное учреждение "Лицей № 130 им. акад. М. А. Лаврентьева", Новосибирск (05.04.2022)

Одной из задач гуманитарных исследований, для решения которых успешно применяются математические методы, является автоматизированный анализ поэтического текста, в процессе которого требуется учитывать строение строки, рифму, размер, а также фонетическое звучание как основу стихосложения. Программное приложение для автоматизации комплексного анализа поэтических текстов разработано в ФИЦ ИВТ: исследования, дополняющие и оптимизирующие данную программную систему, продолжаются. Автоматизированное статистическое исследование транскрипции — один из важных модулей фонетического анализа текста. Программная система выполняет фонетический разбор, однако необходим также модуль анализа статистики. Таким образом, сформулирована задача данного исследования: разработка и реализация приложения для статистического исследования транскрипции русских поэтических текстов.

Разработка и реализация программного модуля построения конкордансов для системы комплексного анализа русских поэтических текстов

Н. А. Шашок

Федеральный исследовательский центр информационных и вычислительных технологий, Новосибирск (05.04.2022)

Конкорданс — расширенный словарь языка поэта — учитывает применение всех возможных словоформ в доступном корпусе произведений автора. Составление конкорданса — задача трудоемкая при большом объеме корпуса текстов, она может быть упрощена с помощью автоматизации этого процесса. В докладе представлен программный модуль для построения конкордансов и работы с ними. Модуль разработан и реализован в рамках системы комплексного анализа поэтических текстов.

Методы генерации лексико-синтаксических паттернов на основе онтологии для извлечения информации о научной деятельности

К. А. Овчинникова

Новосибирский государственный университет, Новосибирск (12.04.2022)

Рассмотрен подход к автоматической генерации лексико-синтаксических паттернов (ЛСП) онтологического проектирования, которые используются для извлечения информации о научной деятельности. Лексико-синтаксические паттерны — это структурные образцы языковых конструкций, которые отображают их лексические и поверхностные синтаксические свойства и описывают объекты предметной области. ЛСП строятся на основе:

– знаний о предметной области научной деятельности, представленных в онтологии научного знания;
– корпуса научных публикаций из разных областей знаний;
– вопросов оценки компетентности.

Реализация алгоритма автоматической генерации ЛСП выполнена на языке Python.

Дополненная внимательная GNN для извлечения связей

Цюаньюань Ван
Новосибирский государственный университет, Новосибирск (12.04.2022)

Одна из проблем работы графовых нейронных сетей состоит в том, что использование матрицы внимания исключительно для замены матрицы смежности дерева зависимостей может постепенно отклониться от исходной синтаксической структуры. В представленной работе в механизм внимания были внесены некоторые изменения с помощью синтаксических деревьев зависимостей. Кроме того, графовая нейронная сеть GNN объединена с представлением модели BERT, чтобы полностью использовать преимущества обеих методик.

Кластеризация русских поэтических текстов с использованием лексико-тематических характеристик

В. Суворов
Новосибирский государственный университет, Новосибирск (19.04.2022)

Доклад посвящен задаче кластеризации русских поэтических текстов на примере корпуса стихотворений А. С. Пушкина. Рассмотрена иерархическая кластеризация по методу одиночной и полной связи, а также представлен алгоритм FRiS-Tax. Для введения метрики используются группы признаков различных уровней модели информации: метроритмические характеристики, пунктуация, лексика и т. д.

Технология построения информационной системы поддержки научных исследований на основе мультиязычного тезауруса

(по материалам кандидатской диссертации)

О. А. Федотова
Государственная публичная научно-техническая библиотека СО РАН, Новосибирск (26.04.2021)

Доклад посвящен разработке модели информационной системы и созданию технологии построения информационного научно-образовательного ресурса на основе семантического тезауруса.

Разработка и реализация модулей системы хранения и комплексного анализа поэтических текстов

Н. А. Шашок
Федеральный исследовательский центр информационных и вычислительных технологий, Новосибирск (17.05.2022)

Представлена реализация некоторых модулей системы автоматизированного комплексного анализа русских поэтических текстов, созданной в ФИЦ ИВТ: составления

конкордансов, метроритмических справочников, словарей языка поэтов, администрирования, статистического анализа транскрипции поэтических текстов, а также изменения в реализации модуля поиска по корпусу текстов.

Определение читабельности предложений на основе синтаксических деревьев

И. А. Смаль
Новосибирский государственный университет, Новосибирск (24.05.2021)

Проблема оценки читабельности — сложности понимания текстов — актуальна, поскольку результаты могут применяться в таких областях, как здравоохранение, образование, маркетинг и др. Решением проблемы занимаются с начала 20-го века, и несмотря на то, что в наше время инструменты для анализа текста, а также построения различных классификаторов и регрессоров достаточно развиты, самыми популярными решениями остаются классические формулы, разработанные в конце 90-х годов.

В докладе представлены результаты построения моделей для оценки читабельности предложений на основе машинного обучения, приведен анализ значимости различных свойств синтаксических деревьев в построенных моделях.

Быстрая адаптация компонентных систем распознавания речи

Д. В. Гребенкин
Новосибирский государственный университет, Новосибирск (31.05.2021)

Цель работы — оценка преимуществ и недостатков различных методов быстрой адаптации компонентных систем распознавания речи на уровне языковых моделей. Скорость адаптации является ключевым моментом для специалистов в области обработки естественного языка или искусственного интеллекта, поскольку она позволяет сделать модели распознавания речи широкого профиля более универсальными. В работе проведен анализ разработки и эффективности применения различных адаптированных модификаций модели распознавания речи vosk-model-ru-0.22 на аудиозаписях, содержащих лексику из новой для модели предметной области, с точки зрения улучшения качества распознавания и количества используемых вычислительных ресурсов.

Становление, развитие и применение количественных методов в анализе русских поэтических текстов. Специализированные системы сбора, хранения и анализа результатов лингвистических исследований

О. Ю. Кожемякина
Федеральный исследовательский центр информационных и вычислительных технологий, Новосибирск (11.10.2022)

Доклад посвящен вопросам применения количественных методов в исследованиях русской поэзии. Рассмотрены существующие системы лингвистического анализа для некоторых языков, а также алгоритмы и методы, лежащие в их основе. Представлены современные исследования, использующие наследие классических методов и получившие благодаря развитию информационных технологий новые векторы продолжения исследований.

Архитектура и ролевые модели модуля конкордансов системы комплексного анализа поэтических текстов

Н. А. Шашок
Федеральный исследовательский центр информационных и вычислительных технологий, Новосибирск (18.10.2022)

Для автоматизации построения конкордансов и словарей языка поэтов в системе комплексного анализа поэтических текстов разработан и реализован специализированный модуль. В докладе описаны общая архитектура модуля, его ролевые модели и реализация алгоритмов взаимодействия моделей в задаче контроля и подтверждения данных.

Применение теории риторических структур к анализу песенной поэзии

А. В. Зоркальцев
Новосибирский государственный университет, Новосибирск (01.11.2022)

Предложены возможные стилеметрические (стилеметрический вес отношения) и сложностные (коэффициент вытянутости и коэффициент имплицитности) меры, которые можно построить на основе деревьев риторических структур. Эти меры вычислены для двух почти равномощных корпусов — текстов песен Юрия Шевчука и Дмитрия Мозжухина. Намечены перспективы работы, в том числе связанные с частичной автоматизацией процесса построения дерева.

Анализ применимости синтаксических характеристик при оценке сложности текстов учебников методами машинного обучения

И. А. Смаль
Новосибирский государственный университет, Новосибирск (08.11.2022)

В докладе представлены результаты построения моделей для оценки сложности текстов учебников по выборкам различных размеров. Построенные модели работают исключительно с синтаксическими признаками текстов, почти не затрагивая их лексические особенности. Приведены результаты анализа значимости признаков различными способами.

Исследование эффективности слоев “плотной” ассоциативной памяти в нейросетевых алгоритмах распознавания речи типа wav2vec2

Д. В. Гребенкин
Новосибирский государственный университет, Новосибирск (15.11.2022)

Современные нейросетевые модели на базе сквозного (end-to-end) подхода являются основой многих прикладных систем компьютерной лингвистики. Не стало исключением и распознавание речи — алгоритмы типа wav2vec2 позволяют создавать более эффективные решения с помощью переноса обучения в глубоких нейросетях, предобученных решению некоторых общих задач, на специализированные задачи в рамках конкретного языка. “Сквозные” нейросетевые алгоритмы чаще всего являются более производительными, требуют меньше вычислительных ресурсов, и их размер может быть уменьшен с помощью методов квантизации или прунинга. Тем не менее известный с 2017 г. механизм внимания, применяемый в большинстве таких нейросетевых алгоритмов, при всех

своих достоинствах не лишен и недостатков. В работе рассматривается возможность замены механизма многоголовочного внимания, используемого в нейросетевом алгоритме распознавания речи типа wav2vec2, на современные сети Хопфилда. Авторами экспериментально проверяется гипотеза о том, что концепция "плотной" ассоциативной памяти лучше подходит для решения задач восстановления пропущенных фрагментов звукового сигнала и преобразования этого сигнала в слова естественного языка, чем многоголовочное внимание. Также проводится попытка теоретического объяснения связи между ассоциативной памятью и языковым контекстом.

Разработка алгоритма анализа синтаксиса текстов узбекского языка

Б. Б. Ибрагимов

Новосибирский государственный университет, Новосибирск (22.11.2022)

Исследование посвящено разработке и реализации программного приложения для проверки синтаксиса текстов на узбекском языке. Зачастую можно наблюдать, как пользователи, набирающие текст на узбекском языке, сталкиваются с проблемой составления и соединения слов в связный текст. К сожалению, на сегодня отсутствуют какие-либо программные решения по проверке корректности объединения слов в предложения. И эту задачу призвано решить создаваемое программное приложение, использующее разработанный алгоритм, основанный на формализации грамматических и синтаксических правил узбекского языка. В перспективе приложение можно будет внедрить как отдельный модуль в такие популярные офисные программные обеспечения, как OpenOffice.org Writer и Microsoft Office Word.

Разработка алгоритмов морфологического анализа словоформ каракалпакского языка

Р. М. Абдуллаев

Новосибирский государственный университет, Новосибирск (22.11.2022)

В Республике Каракалпакстан (Узбекистан) наблюдается активное развитие интернета. Как и в других национальных секторах, самая распространенная форма представления текстовой информации на каракалпакском языке — слабоструктурированные документы, работа с которыми предполагает наличие надежных алгоритмов анализа текста, в том числе его лексических характеристик. Намечены создание и реализация программного приложения для морфологического анализа словоформ каракалпакского языка на основе разработанных ранее алгоритмов. Отметим, что письменность на этом языке, в отличие, например, от узбекского и казахского, возникла уже в советское время, что обусловило сильное влияние на него русского языка (не только на лексику, но и на грамматику) и привело к появлению ряда морфологических особенностей, совершенно не характерных для других тюркских языков.

Разработка веб-приложения для наукометрического анализа университетов из стран СНГ

Б. Р. Саидов

Новосибирский государственный университет, Новосибирск (29.11.2022)

Многие университеты из стран СНГ ориентированы на интенсивное участие в мировой научной повестке и развитие собственной исследовательской базы. При этом они

имеют тесные связи с российскими вузами и научными институтами, нередко выступают партнерами в различных исследовательских проектах и оказывают заметное влияние на российскую науку. Сегодня многое сделано для определения рейтинга вузов в контексте научных исследований, публикационной активности и научной работы в целом. Примером могут быть такие международные базы цитирования, как Scopus, Web of Science, PubMed. Однако на этих ресурсах возможности пользователя ограничены, например, нельзя посмотреть статистику по квартилям источников, где были опубликованы научные труды, или вывести топ университетов по региону (как по одной стране, так и по региону — СНГ). Поэтому приобретает актуальность разработка подобного специализированного сервиса для упрощенного мониторинга и анализа научно-исследовательской активности вузов СНГ.

Алгоритмы автоматизированного определения тематической направленности русских поэтических текстов на основании лексических признаков

Н. С. Каминский

Новосибирский государственный университет, Новосибирск (29.11.2022)

Одна из задач автоматизации комплексного анализа поэтических текстов — разработка алгоритмов, выявляющих тематическую направленность стихотворных произведений. В рамках такой задачи подход с использованием методов машинного обучения, прежде всего нейросетей, представляется наиболее перспективным.

В работе уточнен корпус стихотворений, включающих тему “времена года”, авторства ряда русских поэтов (А. Фета, А. Пушкина, К. Бальмонта, С. Есенина и др.). Эксперименты по использованию уточненного корпуса текстов для обучения нейросетей с целью классификации пейзажной лирики по тематическому признаку “время года” будут продолжены.

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 Наталья Александровна Шашок
e-mail: n.shashok@alumni.nsu.ru
Интерактивная заявка доклада:
http://www.ict.nsc.ru/ru/education/seminar/seminar-page-lingv