МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

Разработка и реализация программного обеспечения, предназначенного для автоматизации процесса обработки и анализа результатов лингвистического ассоциативного эксперимента и поиска семантико-понятийных универсалий

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студентки 2 курса 273 группы
направления 02.04.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Кульгиной Анастасии Алексеевны

Научный руководитель,
к.ф.-м.н., доцент _____ М. В. Огнева

Зав. кафедрой ИиП,
к.ф.-м.н., доцент _____ М. В. Огнева

Саратов 2021

ВВЕДЕНИЕ

Концептуализмов ровно столько,
сколько людей, себя к ним причисляющих,
и каждый это по-своему интерпретирует.
Макс Фрай
(цитируется Л. С. Рубинштейн)

Актуальность темы и значимость магистерской работы обуславливаются тем фактом, что при проведении ассоциативного эксперимента исследователь вынужден совершать вручную множество рутинных действий, требующих огромных время- и трудозатрат, взаимодействуя при этом с огромными объёмами данных (считается [1: 247], что ассоциативный материал идеально репрезентативен, если число реакций на каждый стимул составляет не менее пятисот). Поскольку никаких инструментов автоматизации в области ассоциативной лингвистики на настоящий момент не существует, появление специализированного программного обеспечения[1] позволит сократить количество рутинной работы и необходимое для её выполнения время, повысить степень объективности полученных результатов, снизить влияние человеческого фактора на проводимые эксперименты, а также увеличить охват привлекаемого лингвистического материала.

Цель магистерской работы – разработка и реализация веб-приложения, которое позволит учёным-психолингвистам проводить обработку и анализ ассоциативных данных автоматически.

Поставленная цель определила **следующие задачи**:

1. рассмотреть существующие алгоритмы и методы обработки естественного языка;
2. сделать обзор существующего лингвистического ПО;

[1] Здесь и далее: ПО.

3. сделать краткий обзор психолингвистики как научной области и ассоциативной лингвистики как одного из её направлений;
4. сформировать датасет, содержащий данные ассоциативных экспериментов;
5. выполнить предобработку данных, а также обеспечить пополнение датасета;
6. разработать и протестировать различные алгоритмы для автоматической фреймовой классификации и выбрать среди них наиболее точный;
7. рассмотреть и реализовать в составе веб-приложения статистические и лингвистические методы работы с ассоциативным материалом;
8. провести бета-тестирование с привлечением целевой аудитории разрабатываемого продукта.

Методологические основы компьютерной лингвистики, психолингвистики и ассоциативной лингвистики представлены в работах Ю. Н. Караулова [1], А. Вежбицкой [2], Е. И. Большаковой и др. [3], И. С. Николаевой и др. [4], И. Н. Горелова и К. Ф. Седова [5], А. Зализняк [6], М. Минского [7], В. Е. Гольдина и А. П. Сдобновой [8], Г. Я. Мартыненко и Г. М. Мартинович [9], Е. И. Горошко [10].

Структура и объём работы. Магистерская работа состоит из введения, трёх разделов, заключения, списка использованных источников и трёх приложений. Общий объём работы – 102 страницы, из них 69 страниц – основное содержание, включая 25 рисунков и 4 таблицы, список использованных источников – 51 наименование.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Автоматическая обработка естественного языка» посвящён краткому обзору компьютерной лингвистики как научной области. Рассматриваются различные мнения об истоках данного научного направления,

а также различные определения. В качестве основной в работе принимается следующая трактовка: компьютерная лингвистика – это дисциплина, которая изучает закономерности распределения текстовой информации, проблемы, принципы, методы и алгоритмы разработки лингвистического программного и аппаратного обеспечения.

Кроме того, в данном разделе приводится классификация алгоритмов для автоматической обработки естественного языка, предложенная В. А. Яцко, обсуждаются различные способы векторного представления слов, а также предлагается краткий обзор лингвистического ПО, применяемого на различных языковых уровнях.

Второй раздел «Психолингвистика: краткий обзор» посвящён обсуждению основных понятий психолингвистики и ассоциативной лингвистики, важных в контексте разработанного ПО. Были даны следующие определения:

- психолингвистика – научная область, изучающая особенности порождения, понимания, функционирования и развития речи;
- языковое сознание – внутренний процесс планирования и регуляции внешней деятельности при помощи языковых знаков;
- языковая картина мира – совокупность представлений о мире, сложившаяся в сознании языкового коллектива и отражённая в языке;
- образ сознания – мысленный конструкт, интегрирующий знания, полученные субъектом в ходе речевого общения, а также знания, полученные в результате переработки информации, поступающей от органов чувств в предметной деятельности;
- свободный ассоциативный эксперимент – эксперимент, в ходе которого испытуемому предъявляется список слов-стимулов, на каждое из которых ему предлагается ответить первым пришедшим в голову словом или словосочетанием;

- фрейм – структура данных для представления стереотипной ситуации.

Кроме того, были перечислены фамилии учёных, оказавших наибольшее влияние на психолингвистику как за рубежом (М. Минский, Ч. Э. Осгуд и др.), так и в России (В. Е. Гольдин, Ю. Н. Караулов и др.).

Третий раздел «Реализация приложения» посвящён рассмотрению вопросов, связанных непосредственно с реализацией веб-приложения *Association Analyzer*.

В **подразделе 3.1** кратко освещается полный перечень используемых технологий (Python и Django для серверной части, TypeScript, React и MobX для клиентской части) и их преимуществ перед другими аналогичными решениями.

В **подразделе 3.2** описан процесс сбора и подготовки языкового материала, составившего словарную базу *Association Analyzer* и послужившего обучающим материалом для алгоритма автоматической фреймовой классификации. Данные, полученные от студентов, выпускников и учёных Института филологии и журналистики СГУ им. Н. Г. Чернышевского, работающих и работавших в рамках научно-исследовательского семинара, посвящённого исследованиям языковой личности и языковой картины мира и реализуемого кафедрой теории, истории языка и прикладной лингвистики, прошли несколько стадий подготовки и предобработки.

Сначала было выполнено ручное сведение всего массива ассоциативных пар «стимул – реакция» в единый файл Microsoft Excel, необходимое ввиду того, что материалы были получены в абсолютно различных форматах, начиная от таблиц, оформленных в Microsoft Word, и заканчивая текстовыми приложениями к выпускным квалификационным работам и выдержками из отдельных их глав.

Затем был произведён анализ предоставленных психолингвистами фреймовых структур (общее количество наименований слотов до выполнения данного шага – 101) с целью вывода некой единой системы классификации, не слишком детализированной, не слишком общей и при этом способной отразить структуру потенциально любого образа сознания. Результатом данной

процедуры стала фреймовая структура, включающая в себя следующие классы: «Атрибуты», «Глагольные реакции», «Дефиниция», «Диалогические реакции», «Другое», «Культурный прецедент», «Локализация в пространстве», «Описание», «Оппозиты», «Причинно-следственный компонент», «Реакции-синтагмы», «Рифма», «Участник событий», «Чувства и эмоции».

Далее было написано несколько функций для программной очистки датасета от повторяющихся пар «стимул – реакция», возникших ввиду того, что часть исследований была проведена в сопоставительном ключе.

Наконец все стимулы и реакции были размечены при помощи Python-библиотеки PyMorphy2. Для реакций были выделены следующие ключевые признаки: часть речи, падеж, лицо, количество слов, одушевлённость, вовлечённость говорящего, наклонение, полнота/краткость, инфинитивность, мера семантической близости к стимулу, принадлежность к классу имён собственных. Для стимулов ключевыми признаками являются: количество слов, часть речи, одушевлённость, падеж, вовлечённость говорящего, наклонение, лицо, принадлежность к классу имён собственных, глагольный вид, грамматический род, число, время, категория переходности, категория залога.

В **подразделе 3.3** обсуждается техническая сторона реализации программного продукта.

На первом шаге был реализован механизм регистрации и авторизации пользователей, призванный упростить разграничение доступа к различным частям *Association Analyzer* и хранение результатов действий конкретных пользователей. Помещённые в **пункте 3.3.1** рисунки демонстрируют формы регистрации и авторизации, а также процесс валидации, происходящий на разных этапах взаимодействия пользователя с системой.

На втором шаге был реализован поисковый модуль, включающий в себя две основных стратегии взаимодействия: работа с ассоциативным словарём и работа с отдельной статьёй ассоциативного словаря. Так, *Association Analyzer* предлагает пользователю следующие сценарии: составление прямого или обратного ассоциативного словаря на основе базы ассоциативных данных,

составление прямого или обратного ассоциативного словаря на основе загруженных в систему пользовательских данных, составление статьи прямого или обратного ассоциативного словаря на основе базы ассоциативных данных, составление статьи прямого или обратного ассоциативного словаря на основе загруженных в систему пользовательских данных. В рамках работы с отдельной статьёй ассоциативного словаря был реализован подсчёт следующих лингвостатистических метрик: выделение наиболее и наименее частотных лексем, вычисление главных ассоциатов, количество нулевых реакций, количество лексем-эхолалий, медиана по рангу, первая и третья квартили, дисперсия, энтропия, ассоциативная сила слова, коэффициент лексического разнообразия, коэффициент лексического богатства, мера упорядоченности, мера стереотипности реакций, составление частеречной сводки. Для статьи прямого ассоциативного словаря доступен подсчёт всех перечисленных метрик, в то время как для статьи обратного ассоциативного словаря – лишь тех, наименования которых подчёркнуты одной линией. Дополнительно для каждой статьи как прямого, так и обратного словаря подсчитываются общее число входящих в неё слов и количество среди них уникальных вхождений. Помещённые в **пункте 3.3.2** рисунки демонстрируют взаимодействие пользователя с поисковым модулем.

На третьем шаге было реализовано выполнение автоматической фреймовой классификации ассоциативных реакций. В **пункте 3.3.3** подробно описывается процесс обучения различных алгоритмов, в результате которого для интеграции с *Association Analyzer* был выбран алгоритм градиентного бустинга, представленный в Python-библиотеке `Scikit-learn`. С его помощью удалось добиться 60% точности при классификации на 14 классов и 85.5% точности при классификации на 6 классов. Данная процедура трактуется в рамках системы как метрика, поэтому доступ к ней возможен из поискового модуля в том случае, если пользователь выбрал режим составления словарной статьи прямого ассоциативного словаря.

Наконец на четвёртом шаге было проведено закрытое бета-тестирование созданного программного продукта. В бета-тестировании приняли участие три молодых специалиста в области языкознания, обучающиеся по направлениям бакалавриата и магистратуры в следующих высших учебных заведениях: Саратовский национальный исследовательский государственный университет имени Н. Г. Чернышевского, Московский информационно-технологический университет МАСИ, Санкт-Петербургский государственный университет. В **пункте 3.3.4** подробно описывается процесс тестирования, а также приводятся отзывы бета-тестеров, собранные при помощи анкеты, размещённой на платформе Google Формы. По итогам тестирования *Association Analyzer* производит на потенциальных пользователей положительное впечатление и видится им интересным проектом, за развитием которого они планируют наблюдать в дальнейшем. Вместе с тем было высказано несколько незначительных замечаний и пожеланий, которые уже внесены в бэклог проекта для последующего исправления и/или реализации.

ЗАКЛЮЧЕНИЕ

В настоящей выпускной квалификационной работе магистра рассматривается процесс разработки и реализации ПО, предназначенного для автоматизации процесса проведения лингвистического ассоциативного эксперимента, обработки и анализа его результатов, а также работы с большим объёмом языкового материала, представленного в ассоциативном словаре.

В ходе подготовительных работ были пройдены следующие этапы:

- обзор существующего лингвистического ПО, применимого на различных уровнях системы языка,
- сбор ассоциативного датасета при активном содействии учёных и студентов кафедры теории, истории языка и прикладной лингвистики Института филологии и журналистики СГУ им. Н. Г. Чернышевского,

- ручная подготовка собранных данных к дальнейшей автоматической предобработке,
- анализ предоставленных учёными-психолингвистами вариантов фреймовых структур различных образов сознания и вывод на их основе единой системы классификации, заложенной в дальнейшем в классификационный алгоритм,
- создание признакового пространства стимулов и реакций для обучения классификатора,
- анализ работы различных классификационных алгоритмов машинного обучения с последующим выбором для решения задачи алгоритма градиентного бустинга, представленного в Python-библиотеке `Scikit-learn`, с помощью которого удалось добиться точности в 60% при обучении на полном объёме данных и в 85,5% при обучении на подвыборке основного датасета, содержащей лишь объекты наиболее чётко дифференцируемых классов.

В ходе процесса разработки были реализованы:

- пользовательский интерфейс,
- серверная часть приложения,
- регистрация и авторизация пользователей,
- поиск по прямому и обратному ассоциативному словарю с возможностью загрузки в систему пользовательских данных,
- поиск по отдельным статьям прямого и обратного ассоциативных словарей с возможностью загрузки в систему пользовательских данных,
- подсчёт четырнадцати различных лингвостатистических метрик,
- частеречная разметка стимулов и реакций,
- два варианта автоматической фреймовой классификации – на 6 (при помощи алгоритма, обученного на подвыборке данных) и на 14 классов (при помощи алгоритма, обученного на полном объёме данных).

Наконец, в ходе работы было проведено бета-тестирование разработанного решения с привлечением специалистов в области языкознания.

Результаты работы были частично представлены на следующих научных мероприятиях:

1. студенческая научная конференция «Компьютерные науки и информационные технологии» 2020 г. (22.04.2020, ФКНиИТ СГУ, тема доклада: «Компьютерная психолингвистика и автоматический анализ вербальных ассоциаций»),
2. XI научно-практическая конференция *Presenting Academic Achievements to the World* (03.06.2020, СГУ; по итогам участия доклад *Computational processing of verbal associations: web app project* занял 2 место в секции *Computer Science & Economics*, а также в сборнике трудов конференции была опубликована одноимённая статья),
3. всероссийская конференция молодых учёных «Филология и журналистика в XXI веке», посвящённая 75-летию Победы в Великой Отечественной войне (заочно; по итогу участия в сборнике трудов конференции была опубликована статья «Автоматизация ассоциативного исследования: разработка проекта веб-приложения»),
4. кафедральный этап студенческой научной конференции «Компьютерные науки и информационные технологии» 2021 г. (23.04.2021, ФКНиИТ СГУ; по итогам участия доклад на тему «Автоматическая разметка и анализ вербальных ассоциаций» был рекомендован к участию в следующем этапе конференции),
5. факультетский этап студенческий научной конференции «Компьютерные науки и информационные технологии» 2021 г. (30.04.2021, ФКНиИТ СГУ; по итогу участия доклад на тему «Автоматическая разметка и анализ вербальных ассоциаций» занял первое место и был рекомендован к участию в следующем

этапе конференции, а также к публикации тезисов в сборнике трудов конференции),

6. итоговая студенческая научная конференция СГУ 2021 г. (20.05.2021, СГУ; доклад на тему «Автоматическая разметка и анализ вербальных ассоциаций»).

Настоящая работа имеет потенциал для дальнейшего развития. Так, в ходе бета-тестирования был получен ряд замечаний и пожеланий от пользователей, которые обязательно будут учтены в следующих версиях. Кроме того, представляется необходимым реализовать в дальнейшем ещё ряд функций, направленных на упрощение работы исследователя. Так, есть отдельные группы испытуемых (например, школьники различного возраста), в работе с которыми предпочтение отдаётся чаще очному проведению эксперимента, чем прохождению онлайн-опроса. В связи с этим перспективным представляется добавление в *Association Analyzer* модуля автоматического распознавания рукописного текста, который помог бы облегчить и ускорить оцифровку собранных ответов. Бывают также и случаи, когда исследователю необходимо привлечь как можно больше материалов на различных языках, с которыми он при этом не всегда знаком достаточно хорошо, чтобы обойтись без посторонней помощи, в связи с чем представляется перспективным добавить также модуль автоматического перевода.

Разработанное веб-приложение может уже сейчас стать главным инструментом при проведении психолингвистических исследований для учёных по всей стране. Так, оно не только увеличивает объективность полученных результатов, но и существенно упрощает взаимодействие с ассоциативным материалом и позволяет делегировать большую часть рутинной работы компьютеру, освобождая время исследователя для проверки гипотез и проведения новых экспериментов, ещё более масштабных, чем прежде.

Основные источники информации:

1. Караулов, Ю. Н. Ассоциативная грамматика русского языка. / Ю. Н. Караулов // Москва: Издательство ЛКИ, 2010. 328 с.
2. Вежбицкая, А. Понимание культур через посредство ключевых слов. / А. Вежбицкая // М.: Языки славянской культуры, 2001. 288 с.
3. Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие. / Е. И. Большакова, К. В. Воронцов, Н. Э. Ефремова, Э. С. Клышинский, Н. В. Лукашевич, А. С. Сапин. // Москва: Изд-во НИУ ВШЭ, 2017. 269 с.
4. Прикладная и компьютерная лингвистика. / Под ред. И. С. Николаева, О. В. Митрениной, Т. М. Ландо. Москва: URSS, 2016. 320 с.
5. Горелов, И. Н. Основы психолингвистики. / И. Н. Горелов, К. Ф. Седов. - М.: Издательство «Лабиринт», 2001. 304 с.
6. Зализняк, А. А. Языковая картина мира [Электронный ресурс] / А. А. Зализняк // Универсальная научно-популярная онлайн-энциклопедия «Кругосвет» [Электронный ресурс]. - URL: http://www.krugosvet.ru/enc/gumanitarnye_nauki/lingvistika/YAZIKOVAYA_KARTINA_MIRA.html (дата обращения: 04.12.2019). - Загл. с экрана. - Яз. рус.
7. Minsky, M. A. Framework for Representing Knowledge / M. A. Minsky // Artificial Intelligence Laboratory. Massachusetts Institute of Technology, 1974. P. 81.
8. Гольдин, В. Е., Сдобнова, А. П. Русская ассоциативная лексикография. Саратов: Научная книга, 2008. 77 с.
9. Мартыненко, Г. Я., Мартинович, Г. А. Многопараметрический статистический анализ результатов ассоциативного эксперимента. Изд-во С.-Петерб. ун-та, 2003. 28 с.
10. Горошко, Е. И. Качественные методы анализа данных ассоциативного эксперимента [Электронный ресурс] // Интегративная модель свободного ассоциативного эксперимента. - URL: http://www.textology.ru/article.aspx?aId=95 (дата обращения: 14.12.2020). - Загл. с экрана. - Яз. рус.