ВЫЧИСЛИТЕЛЬНАЯ ОБРАБОТКА КАЗАХСКОГО ЯЗЫКА

СБОРНИК НАУЧНЫХ ТРУДОВ

ҚАЗАҚ УНИВЕРСИТЕТІ
БАСПА ҮЙІ

КАЗАХСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ имени АЛЬ-ФАРАБИ

ВЫЧИСЛИТЕЛЬНАЯ ОБРАБОТКА КАЗАХСКОГО ЯЗЫКА

Сборник научных трудов

Алматы
«Қазақ университеті»
2020

УДК 811.512.122
ББК 81.2Қаз-923
В 27

Рекомендовано к изданию Ученым советом факультета информационных технологий (протокол № 13 от 30 июня 2020 г.)

Сборник выполнен в рамках проекта грантового финансирования научных исследований МОН РК AP05132950 «Разработка информационно-аналитической поисковой системы данных на казахском языке»

Рецензенты:
PhD, доцент, кандидат физико-математических наук ***К.С. Дуйсебекова***
кандидат технических наук ***Л.С. Копболсын***

Под редакцией
PhD Рахимовой Д.Р.

В 27 **Вычислительная** обработка казахского языка: сборник научных трудов / под редакцией Рахимовой Д.Р. – Алматы: Қазақ университеті, 2020. – 147 с.

ISBN 978-601-04-4698-4

В данном сборнике представлены научные разработки в области обработки казахского языка.

Предназначен для преподавателей, научных сотрудников, магистров и студентов факультетов информационных технологий и филологий.

УДК 811.512.122
ББК 81.2Қаз-923

ISBN 978-601-04-4698-4

СОДЕРЖАНИЕ

ВВЕДЕНИЕ 4

Глава 1. РАЗРАБОТКА ОНОТОЛОГИЧЕСКОЙ МОДЕЛИ ГРАММАТИКИ КАЗАХСКОГО ЯЗЫКА (Шарипбай А.А., Муканова А.С., Ергеш Б.Ж., Разахова Б.Ш., Елибаева Г.К.) 6

Глава 2. К ВОПРОСУ О РАЗРАБОТКЕ КОРПУСА КАЗАХСКОГО ЯЗЫКА (Мадиева Г.Б., Бектемирова С.Б., Мамбетова М.К.) 34

Глава 3. РАЗРАБОТКА МЕДИА-КОРПУСА КАЗАХСКОГО ЯЗЫКА (Мансурова М.Е., Мадиева Г.Б., Қадырбек Н.Қ., Қыргызбаева М.Е.) 48

Глава 4. РАЗРАБОТКА АЛГОРИТМОВ ИЗВЛЕЧЕНИЯ КЛЮЧЕВЫХ СЛОВ И СЕМАНТИЧЕСКОГО АНАЛИЗА ДАННЫХ НА КАЗАХСКОМ ЯЗЫКЕ (Рахимова Д., Турганбаева А.О., Жуманов Ж.М.) 63

Глава 5. МОДЕЛИ И МЕТОДЫ СЕНТИМЕНТ АНАЛИЗА ТЕКСТОВ НА КАЗАХСКОМ ЯЗЫКЕ (Ергеш Б.Ж., Шарипбай А.А., Бекманова Г.Т.) 86

Глава 6. МЕТОД ИДЕНТИФИКАЦИИ КРИМИНАЛЬНОГО ЗНАЧЕНИЯ ТЕКСТОВ НА КАЗАХСКОМ ЯЗЫКЕ, БАЗИРУЮЩИЙСЯ НА VSM (Мамырбаев О.Ж., Хайрова Н.Ф., Мухсина К.Ж., Колесник А.С.) 104

Глава 7. РАЗРАБОТКА МОДЕЛИ ПОСТ-РЕДАКТИРОВАНИЯ В МАШИННОМ ПЕРЕВОДЕ КАЗАХСКОГО ЯЗЫКА (Рахимова Д.Р., Турарбек А.Т., Пазылхан Н.М.) 121

ВВЕДЕНИЕ

Обработка естественного языка (*Natural Language Processing, NLP*) – общее направление искусственного интеллекта и математической лингвистики. Оно изучает различные прикладные задачи, связанные с информационными технологиями анализа и синтеза естественных языков. Применительно к искусственному интеллекту анализ означает понимание языка, а синтез – генерацию грамотного текста. Решение этих проблем будет означать создание более удобной формы взаимодействия компьютера и человека. Сегодня NLP применяется во многих сферах, в том числе в голосовых помощниках, автоматических переводах текста и фильтрации текста.

В области NLP проводятся исследования и разработки различных задач. Среди этих задач, можно выделить следующие:

– Распознавание текста, речи, синтез речи;

– Морфологический анализ (слова);

– Сбор и хранение лингвистических ресурсов (корпус, словари, тезаурус и др.);

– Синтаксический разбор, токенизацию предложений;

– Извлечение отношений, определение языка, анализ эмоциональной окраски;

– Аннотацию документа, перевод, анализ тематики текста;

– Информационный поиск, машинный перевод и др.

Область обработки естественного языка в мировой науке является достаточно зрелой, разработаны многие формальные модели, методы и алгоритмы. Существуют высокого уровня различные прикладные программные системы по анализу и обработке естественных языков, сбору и хранению лингвистических данных.

Активная интеграция Казахстана в мировое сообщество и увеличивающимся объемом информационных потоков между нашей страной и ее зарубежными партнерами, реальная потреб-

ность для различных слоев населения в информационных технологиях в повседневной жизни возрастает. В данной книге представлены научные исследования и разработки в области обработки казахского языка. Научное издание состоит из семи глав различных работ научных групп из исследовательских институтов, университетов и лабораторий Республики Казахстан. В каждой главе представлены различные научные исследования, разработанные алгоритмы, модели и технологии по различным тематикам обработки казахского языка. Представленные научные направления являются результатами долгих и трудоемких работ. Издание предназначено для преподавателей, научных сотрудников, магистров и студентов факультетов информационных технологий и филологии.