МИНОБРНАУКИ РОССИИ

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

Кафедра теории, истории языка и прикладной лингвистики

Особенности автоматического морфологического анализа русского языка (на примере модуля нормализации текстов)

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 5 курса 511 группы
направления 45.03.01 «Филология», профиля «Отечественная филология»
(Русский язык и литература)
Института филологии и журналистики
Александровой Камилы Александровны

Научный руководитель
к.ф.н. _____ _____ А.И. Буранова

Зав. кафедрой
д.ф.н., профессор _____ _____ О.Ю. Крючкова

Саратов 2023

Цель исследования заключается в разработке и реализации эффективных алгоритмов и подходов для нормализации текста в контексте обработки естественного языка и синтеза речи.

Обозначенная цель предполагает решение следующих задач:

– охарактеризовать компьютерную морфологию как один из этапов обработки естественного языка, включая основные подходы и обзор программ;

– изучить задачу нормализации текста, её место в рамках других задач по обработке естественного языка и задачи синтеза речи;

– рассмотреть проблему нормализации, включая обзор программ и актуальность;

– построить эвристику и выбрать подход к задаче нормализации;

– разработать алгоритмы обработки для нормализации текста;

– реализовать алгоритм нормализации текста.

Объектом исследования являются алгоритмы и методы нормализации текста в контексте обработки естественного языка.

Источниками послужили различные научные и учебные издания. Основной опорой исследования стали учебные пособия «Автоматическая обработка текстов на естественном языке и анализ данных» Е.И. Большаковой и «Прикладная и компьютерная лингвистика» И. С. Николаева (под ред. Т. М. Ландо), а также ресурсы ACL Anthology, IEEE Xplore и другие.

Структура выпускной квалификационной работы бакалавра: работа состоит из введения, двух глав, заключения, списка использованных источников и литературы.

Основное содержание работы. Первая глава «Теоретические аспекты исследования» посвящена основам компьютерной морфологии и задаче нормализации текста в контексте обработки естественного языка и задач синтеза речи. Глава состоит из трех подразделов: «Компьютерная морфология как один из этапов обработки естественного языка. Задачи,

основные подходы, обзор программ», «Задача нормализации текста. Её место в рамках других задач по обработке естественного языка и задачи синтеза речи», «Проблема нормализации: обзор программ, актуальность».

В первом параграфе «Компьютерная морфология как один из этапов обработки естественного языка» рассматриваются задачи и основные подходы к изучению компьютерной морфологии. Это важный этап обработки естественного языка, который включает в себя анализ, понимание и интерпретацию человеческого языка в целях его дальнейшей обработки и использования.

Второй параграф «Задача нормализации текста. Её место в рамках других задач по обработке естественного языка и задачи синтеза речи» посвящен вопросам нормализации текста. Нормализация текста — это процесс преобразования текста в единый стандартный вид, который упрощает последующую обработку и анализ [1]. Этот подраздел также рассматривает место задачи нормализации в контексте других задач обработки естественного языка и задач синтеза речи.

Третий параграф «Проблема нормализации: обзор программ, актуальность» обсуждает актуальность проблемы нормализации и представляет обзор программ, которые используются для решения этой задачи.

В целом, первая глава представляет собой теоретическую основу для дальнейшего исследования и практической реализации задачи нормализации текста.

Во второй главе «Практические аспекты нормализации текста» проводится детальный анализ процесса нормализации текста, рассматривается его функционирование в контексте обработки естественного языка (на материале сопоставления текстовых корпусов до и после проведения процедуры нормализации).

Каждый из выделенных этапов нормализации текста был описан с точки зрения а) методов и инструментов, используемых для выполнения

этого этапа, б) целей и задач, которые решаются на данном этапе, и в) его функционирования.

В целом, вне зависимости от используемого в работе подхода выделяется основной алгоритм, состоящий из нескольких этапов. Каждый этап (методы и инструменты, цели и задачи) был представлен в виде структуры, включающей основные и дополнительные элементы. Критерием построения этой структуры на материале корпуса текстов послужила частота использования определенных методов и инструментов, а также их эффективность в решении задач нормализации текста. В работе были представлены следующие этапы:

1. Преобразование текста в нижний регистр

Этот этап является одним из наиболее важных в процессе нормализации. Он позволяет унифицировать текст и сделать его более удобным для дальнейшей обработки. Также были выявлены случаи, когда преобразование в нижний регистр может привести к потере важной информации, например, при работе с именами собственными: городами, названиями улиц, и т. д.

2. **Удаление знаков препинания**

Этот этап позволяет убрать из текста элементы, которые не несут семантической нагрузки, и сосредоточиться на словах и их последовательности. Однако, как и в случае с преобразованием в нижний регистр, удаление знаков препинания может привести к потере важной информации, например, при работе с сокращениями: например, адресными (“г.”, “ул.”).

3. **Токенизация**

Токенизация текста является ключевым этапом в процессе нормализации. Этот процесс включает разбиение текста на отдельные слова или “токены” [2], [3]. Это позволяет системам обработки естественного языка анализировать и понимать текст на более глубоком уровне, так как каждое слово или токен могут быть проанализированы и обработаны отдельно.

4. **Поиск нестандартных записей и их обработка**

Этот этап рассматривается исключительно в рамках выбранного подхода. Всего в работе рассмотрено 3 основных подхода и соответствующие им алгоритмы поиска и обработки нестандартных записей [4]:

1. ***Подход на основе правил.*** Этот подход подразумевает создание наборов правил и шаблонов для обработки различных типов текстовых элементов, таких как числа, даты, временные метки, аббревиатуры и т. д. Он может включать создание словарей и грамматик, которые определяют преобразования и соответствия между исходными текстами и их произносимыми формами. Подход на основе правил может быть эффективным, но он также может быть трудоёмким и сложным для поддержки, особенно при работе с большими объемами данных или при необходимости обрабатывать новые типы текстовых данных.

2. ***Подход с использованием машинного обучения***. С использованием алгоритмов машинного обучения и больших наборов данных с размеченными примерами нормализации текста можно создать модели, которые обучаются на основе этих данных и могут применять эти навыки для нормализации нового текста. Методы машинного обучения, такие как классификация и регрессия, могут быть использованы для создания эффективных моделей нормализации текста. Машинное обучение может быть более гибким и мощным подходом, но оно также требует больших объемов размеченных данных для обучения и может быть сложным в настройке и интерпретации.

3. ***Подход с использованием глубокого обучения и нейронных сетей.*** Модели глубокого обучения, такие как рекуррентные нейронные сети (RNN), трансформеры и другие архитектуры, способны обрабатывать последовательности данных и могут быть обучены для нормализации текста. Глубокое обучение может обеспечить еще большую гибкость и

мощность, позволяя моделям автоматически изучать сложные паттерны и зависимости в данных. Однако эти модели также могут быть сложными в настройке и требовать значительных вычислительных ресурсов.

В заключении работы подчеркивается важность задачи нормализации текста в области обработки естественного языка и синтеза речи. Для решения этой задачи было использовано несколько подходов, в рамках которых были разработаны различные алгоритмы и методы, включая построение эвристик, которые были рассмотрены в данной работе. Была реализована система нормализации текста, которая может быть использована для улучшения качества обработки естественного языка и синтеза речи.

Важно отметить, что нормализация текста является сложной и многогранной задачей, требующей дальнейшего исследования и разработки. Возможно, в будущем появятся новые методы и подходы, которые позволят еще более эффективно решать эту задачу.

Разрешимость задачи нормализации текста для его последующего озвучивания во многом зависит от качества используемых методов и алгоритмов, а также от сложности и особенностей конкретного языка. Для русского языка существуют различные подходы и инструменты, которые успешно решают эту задачу, в большинстве своем используя машинное обучение.

Список цитируемой литературы:

1. Николаев И.С. Прикладная и компьютерная лингвистика. / И. С. Николаев, О. В. Митренина, Т. М. Ландо. (ред.). - М. : URSS, 2016. - 320 с.

2. Bird S. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. / S. Bird, E. Klein, E. Loper. O'Reilly Media, 2009. - 502 p.

3. Jurafsky D. Speech and Language Processing (2nd Edition). / D. Jurafsky, J. H. Martin. - Prentice Hall, 2008. - 1024 p.

4. Автоматическая обработка текстов на естественном языке и анализ данных : учеб. пособие / Е. И. Большакова [и др.] - М. : Изд-во НИУ ВШЭ, 2017. - 269 с.