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

Кафедра информатики и программирования

АВТОМАТИЧЕСКОЕ ИСПРАВЛЕНИЕ ОШИБОК И ОПЕЧАТОК В РУССКОЯЗЫЧНЫХ ТЕКСТАХ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

Студентки 2 курса 273 группы
направления 02.04.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Маниной Дарьи Романовны

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

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

Саратов 2022

ВВЕДЕНИЕ

Каждый пользователь когда-либо допускал ошибки или опечатки при наборе текста. Отсутствие механизмов, которые способны их исправить, может привести к неприятным ситуациям различной степени. Даже при обычной переписке такие ошибки могут привести к непониманию, а есть предметные области, в которых задача хранения безошибочных документов в электронной форме имеет особую значимость. С появлением компьютеров и других устройств, с распространением различных сервисов и услуг в интернете, резко возрастает количество текстовых документов, а вместе с тем возрастает и острота проблемы обнаружения и исправления в них ошибок.

Порой ошибка или опечатка в документе может признать его недействительным, случается, что текст может приобрести абсолютно иной смысл. Также наличие ошибок в запросе влияет на отклик поисковых систем в интернете, что может привести к выдаче нерелевантных результатов при поисковых запросах, а то и вовсе к их отсутствию.

Задача поиска и исправления ошибок в текстах, является одной из самых непростых в области автоматической обработки и анализа текстов на естественных языках. История попыток решить данную задачу берёт своё начало с середины шестидесятых годов в работах Владимира Иосифовича Левеншейна и Фредерика Дамерау, в которых были предложены метрики, позволяющие определить расстояние между двумя словами. На основе этих метрик подбираются наиболее близкие слова из словаря, которые рассматриваются в качестве кандидатов на замену слова не из словаря.

В языке слова редко используются отдельно и зависят от контекста, в котором они используются, поэтому последовательность правильных слов может образовывать неправильное предложение. Для того, чтобы проводить исправления таких ошибок необходимо использование языковой модели.

В общем случае механизм исправления опечаток и ошибок основывается на двух моделях: модели ошибок или языковой модели. Для контекстно-независимого исправления используется только модель ошибок, а в контекстно-зависимом и модель ошибок, и языковая модель.

Существующие современные системы обработки ошибок и опечаток разделяются на два типа: программы обнаружения опечаток, которые лишь обнаруживают опечатки и ошибки и сообщают о них пользователю, например, автопроверка орфографии в браузерах или в текстовых редакторах, таких как «Microsoft Word» и программы исправления опечаток и ошибок, которые самостоятельно принимают решение об исправлении обнаруженных ими ошибок, например, автоматическое исправление в поисковых запросах «Яндекса» и «Google». Однако, несмотря на действующие системы и различные исследования не существует универсального решения данной проблемы, которое в любом случае дает идеальный результат, поэтому поиск новых, более эффективных решений продолжается.

Обычно программы автоматического исправления используются для небольших запросов, состоящих из одного предложения или даже нескольких слов. В данной работе проводится исследование методов автоматического исправления в тексте (совокупности последовательно расположенных предложений, связанных по смыслу).

Целью данной работы является реализация и исследование методов для решения задачи автоматического поиска и исправления ошибок в русскоязычных текстах на основе совместного и независимого применения языковой модели и модели ошибок.

Данная цель определила следующие **задачи**:

1. Проанализировать современные методы обнаружения и исправления ошибок в русскоязычных текстах.
2. Собрать и предобработать корпус русскоязычных текстов.
3. Провести оценку влияния различных стилей текстов на количество

уникальных слов, биграмм, триграмм и 5-грамм.

4. Сделать выводы о возможности использования набора полученных N-грамм на одном стиле для текстов, относящихся к иным стилям.
5. В качестве модели ошибки рассмотреть расстояние Дамерау-Левенштейна и реализовать его подсчёт.
6. Рассмотреть и реализовать языковую модель на основе N-грамм.
7. Реализовать методы исправления ошибок на основе модели ошибок и модели языка.
8. Провести оценку качества реализованных методов.
9. Сделать выводы.

Методологические основы автоматического исправления ошибок представлены в работах В.И. Левенштейна, Т.О. Шавриной, Фредерика Дамерау, Кристофера Брайанта, Феликса Штальберга.

В теоретической части изложена общая информация об автоматическом исправлении ошибок и опечаток, которое является контекстно-зависимым. Данное исправление основано на модели ошибок и модели языка. В качестве модели ошибок было рассмотрено расстояние Дамерау-Левенштейна, а в качестве модели языка — метод на основе N-грамм. Подробно описан принцип работы данных двух моделей.

В практической части работы описан процесс сбора и предобработки корпуса русскоязычных текстов. Проведена оценка влияния различных стилей текстов на количество уникальных слов, биграмм, триграмм и 5-грамм. В качестве модели ошибки реализовано расстояние Дамерау-Левенштейна, а для модели языка реализована языковая модель на основе N-грамм.

Структура и объём работы. Магистерская работа состоит из введения, пяти разделов, заключения, списка использованных источников и восьми приложений. Общий объем работы – 72 страницы, из них 54 страницы – основное содержание, включая 28 рисунков и 10 таблиц, список использованных источников информации – 27 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Методы обнаружения и исправления опечаток» посвящен обзору методов исправления ошибок и опечаток.

Ранние методики коррекции орфографии не давали высокого качества исправления, в связи с этим, возросла необходимость улучшения вероятностной модели отбора кандидата. Также стало очевидно, что это невозможно без применения контекстных методов, когда для повышения качества исправления ошибок, следует учитывать окружение слова — контекст.

В настоящее время механизм исправления опечаток и ошибок основывается на двух моделях: модели ошибок или языковой модели. Для контекстно-независимого исправления используется только модель ошибок, а в контекстно-зависимом и модель ошибок, и языковая модель.

Второй раздел «Модель ошибок» посвящен описанию модели ошибок.

Модель ошибок позволяет исправлять ошибки в отдельных словах — пропуск, вставка, перестановка и замена букв в словах, ошибки слитно-раздельного написания — пропуск или вставка лишнего пробела между словами.

Использование данной модели заменяет слова с помощью вычисления расстояния между ними посредством попарного сравнения символов. Части слов, которые близки по звучанию, также близки и в письменном виде. Можно охарактеризовать задачу определения сходства слов по произношению путем подсчета расстояния между словами по написанию.

В качестве модели ошибок обычно выступают редакционные расстояния. Наиболее известными редакционными расстояниями являются расстояния Левенштейна и расстояние Дамерау-Левенштейна.

В подразделе «Редакционное предписание» дается определение понятия и пример для двух строк.

В подразделе «Расстояние Левенштейна» дается определение понятия и где оно применяется. Также выводится формула для подсчета расстояния Левенштейна, приводится пример расчета матрицы и нахождения расстояния Левенштейна для двух конкретных строк.

В подразделе «Расстояние Дамерау-Левенштейна» дается определение понятия, выводится формула для подсчета расстояния Дамерау-Левенштейна, приводится пример расчета матрицы и нахождения расстояния Дамерау-Левенштейна для двух конкретных строк.

В подразделе «Мера схожести по Левенштейну» выводится формула для расчета меры для двух строк.

В подразделе «Типы ошибок» проводится разделение на два класса ошибок, встречающихся в русскоязычных текстах.

Третий раздел «Модель языка» посвящен описанию модели языка.

Для работы контекстно-зависимого исправления требуется языковая модель. Любая модель языка начинается с собирания большого корпуса текстов и подсчета встречаемости слов в нем. Задача превращается в задачу вычисления $P(w_1, \ldots, w_m)$, где одно из w_i — слово, в котором исправили опечатку и для которого теперь рассчитываем $P(w)$, а остальные w_i — слова, окружающие исправляемое слово в тексте.

В подразделе «Метод N-грамм» описывается принцип, на котором основывается данный метод, а также приводится пример разбиения предложения на N-граммы.

В подразделе «Языковая модель на основе N-грамм» описывается принцип работы языковой модели, построенной на подсчете частот N-грамм в словах текстов с учетом позиционности N-граммы.

В подразделе «Контекстные методы для русского языка» осуществляется разбор информации, которую можно использовать для ранжирования кандидатов на исправление.

Четвертый раздел «Практическая часть» посвящен описанию процесса сбора корпуса русскоязычных текстов для исследования языка, предобработке и оценке полученного корпуса. Также в разделе описаны методы исправления ошибок на основе модели ошибок и модели языка и определена оценка качества работы данных методов.

В подразделе «Сбор корпуса текстов» описывается процесс сбора публицистических и художественных русскоязычных текстов и дальнейшая предобработка полученного корпуса, состоящего из этих текстов.

В подразделе «Хранение корпуса и языковой модели» рассматривается работа с базой данных SQLite для упорядочивания и хранения языковой модели. Таким образом, взаимодействие с моделью языка происходит через запросы к базе данных. Языковая модель хранит в себе биграммы, триграммы и 5-граммы предложений из текстов корпуса.

В подразделе «Оценка корпуса» проводится оценка собранного корпуса для определения достаточности данных для исследования. При насыщении корпуса одним стилем, можно перейти на другой, для большего разнообразия контекста. Для коррекции текстов определенного стиля в корпусе должен находиться достаточный объем текстов данного стиля.

В подразделе «Выбор кандидата для замены» описываются четыре метода для предугадывания возможного кандидата на замену:

1) с учетом только модели ошибок;
2) с учетом сначала модели языка, а затем модели ошибок;
3) с учетом сначала модели ошибок, а затем модели языка;
4) с учетом модели ошибок и модели языка.

В подразделе «Оценка качества работы методов» определяется каким образом оценивать работу методов исправления, на каких данных и какими метриками.

Пятый раздел «Результаты исследования» обобщает полученные результаты исправления ошибок и опечаток по метрикам «точность», «погрешность» и «корректность» в таблицу.

Благодаря таблице видно, что результат независимого друг от друга использования модели языка и модели ошибок приводит к наиболее эффективному исправлению. Это происходит в следствии того, что две данных модели самостоятельно генерируют кандидатов на замену, не опираясь друг на друга, а в результате выбирается кандидат их объединения, суммарный по характеристикам, которая выдает каждая из моделей.

Также на рисунках, в виде диаграмм, даются оценки нахождения кандидата на исправление в контексте и исправления слов в зависимости от размера N-граммы. Из этих двух оценок можно сделать вывод, что слово наиболее зависимо от предыдущих двух слов до него.

В завершении раздела в таблице представлены результаты сравнения независимого друг от друга использования модели языка и модели ошибок с уже имеющейся системой Яндекс Спеллер, используемой разработчиками для интерактивной проверки орфографии на страницах сайтов.

Из таблицы видно, что Яндекс Спеллер показывает более эффективное исправление ошибок и опечаток. Это связано с тем, что Спеллер − готовый продукт и его языковые модели включают сотни миллионов слов и словосочетаний, что значительно сказывается на обнаружении и исправлении ошибок.

ЗАКЛЮЧЕНИЕ

Обнаружения ошибок в тексте является одним из наиболее трудоемких процессов обработки информации. Вместе с объемом вводимой в ЭВМ информации возрастает и острота проблемы обнаружения и исправления ошибок в ней.

Исправление опечаток разделяется на контекстно-независимое и контекстно-зависимое (где учитывается словарное окружение). В первом случае ошибки исправляются для каждого слова в отдельности, во втором — с учетом контекста. Использование контекстно-зависимого метода позволяет повысить эффективность и корректность исправления опечаток и ошибок.

Основной целью данной работы являлось исследование методов исправления орфографических ошибок и опечаток в русскоязычных текстах с учетом контекста, которые основываются на двух моделях: модель ошибок и языковая модель.

Результаты исследования показали, что при работе с текстами определенных стилей следует использовать корпус текстов того же стиля для заполнения базы данных, поскольку автоматическое исправление ошибок зависимо от контекста.

Независимое использование модели языка и модели ошибок ведет к более точным результатам исправления. Это связано с тем, что каждая модель генерирует свой список кандидатов на замену, опираясь на присущие ей характеристики, а результатом будет являться общий кандидат, принадлежащий обеим моделям и с наивысшей суммарной оценкой по двум моделям. Такой метод позволяет уравновесить влияние на генерацию кандидатов для двух моделей. Таким образом и модель языка, и модель ошибок в равной степени действуют на исправление ошибок и опечаток.

По тематике магистерской работы были представлены доклады:

1. «Методы автоматического исправления ошибок в электронных документах» на XII Всероссийской научно-практической конференции «Информационные технологии в образовании» «ИТО-Саратов-2020»,

Саратов, 30-31 октября 2020 года. Доклад опубликован в материалах конференции.

2. «Методы автоматической коррекции опечаток» на студенческой научной конференции «Компьютерные науки и информационные технологии», Саратов, 29 апреля 2022 год.

Основные источники информации:

1. Левенштейн, В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов / В.И. Левенштейн // Доклады Академий Наук СССР. – 1965. – Т.163, №4. – С.845–848.
2. Damerau, F.J. A technique for computer detection and correction of spelling errors / F.J. Damerau // Communications of the ACM. – 1964. – Vol. 7, №3. – P.171–176.
3. Bryant, C. Language Model Based Grammatical Error Correction without Annotated Training Data / C. Bryant, T. Briscoe // Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications. – 2018. – P. 247–253.
4. Stahlberg, F. Neural Grammatical Error Correction with Finite State Transducers / F. Stahlberg, C. Bryant, B. Byrne // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – 2019. Vol. 1. – P. 4033–4039.
5. Шаврина, Т.О. Моделирование расширенной лемматизации для русского языка на основе морфологического парсера TnT-Russian / Т.О. Шаврина, А. А. Сорокин / Компьютерная лингвистика и интеллектуальные технологии. – 2017. С.10–25.