Анализ экспериментальных данных выбора варианта при нечетком сравнении строк

И. Е. Воронина, e-mail: irina.voronina@gmail.com
Н. А. Экерт, e-mail: ekertn7@gmail.com

Воронежский Государственный Университет

Аннотация. *Рассматривается актуальная проблема выявления опечаток и орфографических ошибок в тексте. Приводятся результаты анализа экспериментальных данных, полученных с целью улучшения ранее представленного алгоритма контекстно-независимого выбора варианта при нечетком сравнении строк.*

Ключевые слова: *нечеткое сравнение строк, расстояние Джаро-Винклера, исправление опечаток, исправление орфографических ошибок.*

Введение

Задача анализа комментариев в социальных и корпоративных сетях не теряет своей актуальности. Одной из важных проблем является корректное представление информации, другими словами, выявление орфографических ошибок, опечаток и т.д. Поэтому в основе решения задачи лежит разработка алгоритма выбора варианта нечеткого сравнения строк.

Задача является объемной и нетривиальной, что требует введения ряда ограничений. Анализу будут подвергаться отзывы, которые оставляют работники организации (предприятия) в корпоративных сетях. Частным случаем является отзывы на изменение версий программного обеспечения. Введенные ограничения позволят избежать использования ненормативной лексики и сленговых выражений. Исследование нацелено на пользователей с достаточным уровнем интеллекта и образования.

На международной научной конференции «Актуальные проблемы прикладной математики, информатики и механики» (Воронеж, 12-14 декабря 2022) авторами была проведена классификация некоторых видов ошибок, совершаемых пользователями в процессе ввода текста. Для анализа использовались следующие виды ошибок:

1. Ошибки слитно-раздельного написания (перечисленныеметоды → перечисленные методы, приве тчеловек → привет человек, м олоко → молоко).

2. Орфографические ошибки (кинот**и**атр → кинот**е**атр).
3. Опечатки (ло**м**ось → ло**с**ось).
4. Дубликаты символов (клавиат**уууууур**а**аааа** → клавиат**ур**а**)**.
5. Ложные нажатия (прип**ео**в → прип**е**в).

2. Метод нечеткого сравнения строк

Методы нечеткого сравнения строк хорошо подходят для решения проблемы нахождения опечаток и орфографических ошибок в тексте [1]. Данные методы предназначены для поиска похожих, но не совпадающих в точности строк. Ранее нами был произведен анализ наиболее известных методов нечеткого сравнения строк, среди которых алгоритм поиска редакционного расстояния Левенштейна, Дамерау-Левенштейна, расстояния Хэмминга и Джаро-Винклера.

Для каждого из методов были разработаны программные средства на языке Python 3 без применения сторонних решений. Оценка скорости работы рассматриваемых методов производилась на строках различной длины от 10 до 100 символов с шагом в 10 символов. Алгоритмы были протестированы с использованием и без использования словаря на предварительно сформированном наборе данных, произведена оценка по следующим критериям: сложность, время выполнения алгоритма без использования и с использованием словаря. Результаты тестирования алгоритмов приведены в табл. 1.

Таблица 1

Результаты тестирования алгоритмов

Алгоритмы	Метрики		
	Сложность	Время выполнения алгоритма без использования словаря, сек	Время выполнения алгоритма с использованием словаря, сек
Расстояние Хэмминга	O(n)	$25{,}9 \cdot 10^{-6}$	$462 \cdot 10^{-3}$
Расстояние Джаро-Винклера	O(n)	$21{,}3 \cdot 10^{-6}$	$486 \cdot 10^{-3}$
Расстояние Левенштейна	$O(n^2)$	$62{,}3 \cdot 10^{-6}$	17
Расстояние Дамерау-Левенштейна	$O(n^2)$	$1{,}78 \cdot 10^{-3}$	43

При сравнении алгоритмов были выявлен ряд отличий. Так, например, алгоритм Хэмминга наделен существенным недостатком – возможностью проводить сравнение на строках только равной длины. Алгоритм поиска редакционного расстояния Левенштейна и Дамерау-Левенштейна более чем в 120 раз уступают алгоритму Джаро-Винклера и Хэмминга по скорости работы на длине строк превышающих 20 символов из-за высокой алгоритмической сложности.

Ключевым критерием при выборе алгоритма являлось время его выполнения. Таким образом, за основу при проведении дальнейшего исследования был выбран алгоритм нечеткого сравнения строк Джаро-Винклера [2–5], дополнительным преимуществом которого является выдача нормированного результата по умолчанию.

В случае наличия множественного выбора при нечетком сравнении строк, предлагается контекстно-независимый алгоритм выбора варианта, позволяющий выбрать наиболее подходящее слово из предложенных вариантов с некоторым коэффициентом уверенности.

Алгоритм разделен на несколько этапов: выбор варианта с учетом опечаток, с учетом орфографических ошибок и на основе экспериментально полученных данных.

3. Выбор варианта с учетом опечаток

Ранее была сформирована гипотеза о распределении весовых коэффициентов в зависимости от удаленности символов на устройстве ввода. Сформирован ряд правил, используемых для формирования словаря распределения символов. Важно подчеркнуть влияние выбранного источника ввода (клавиатура персонального компьютера, виртуальная клавиатура мобильного устройства) на формирование словаря.

В случае использования клавиатуры персонального компьютера в качестве устройства ввода, предполагается, что коэффициент уверенности при выборе соседствующего символа тем больше, чем больше площадь соприкосновения клавиш. В случае если площадь соприкосновения двух и более клавиш одинакова, выбирается тот символ, расстояние до которого от одной из крайних точек на клавиатуре минимально. Крайние точки определяются как точки, максимально удаленные от центра клавиатуры.

В случае использования виртуальной клавиатуры мобильного устройства, предполагается, что инструменты автокоррекции и предиктивного ввода отключены, коэффициент уверенности при выборе соседствующего символа тем больше, чем меньше расстояние от символа до нижней правой крайней точки на клавиатуре.

4. Выбор варианта с учетом орфографических ошибок

Ранее было выдвинуто предположение о распределении весовых коэффициентов с учетом частоты встречаемости орфографических ошибок, составлен словарь распределения символов, представленный в табл. 2.

Таблица 2

Распределение коэффициентов уверенности при выборе символов

Заменяемый символ	Коэффициент уверенности замены (k)			Пример
	3	2	1	
А	О			М**а**л**а**ко → м**о**л**о**ко
Б	П			Оши**п**ка → оши**б**ка
В	Ф			Верё**ф**ка → верё**в**ка
Г	Х			Сне**х** → сне**г**
Д	Т			Бу**т**ка → бу**д**ка
Е	Ё	И	Э	М**и**тель → м**е**тель, к**э**та → к**е**та
Ё	Е	О		Ш**е**пот → ш**ё**пот, ш**е**л → ш**ё**л
Ж	Ш			Эта**ш** → эта**ж**

5. Выбор варианта на основе экспериментально полученных данных

Для реализации этапа выбора варианта на основе экспериментально полученных данных необходимо проведение вычислительного эксперимента с привлечением определенной группы людей [6]. Для формирования группы использовался метод случайного отбора с целью обеспечения репрезентативности и содержательности экспериментальной выборки. Было привлечено 17 человек различного пола в возрасте от 19 до 57 лет.

Для исследования сформирован набор данных, включающий 1000 предложений из известных литературных произведений. Участники эксперимента выполняли скоростной набор текста через предоставленный им графический интерфейс. В качестве устройства ввода использовалась клавиатура персонального компьютера. Предложения выбирались в случайном порядке.

Каждый участник совершал равное количество наборов текста. За время проведения эксперимента было собрано 1700 предложений, включающих 1378 слов, написанных с ошибками. В случае искажения слова при печати, оно служит исходным материалом для пополнения

набора данных. Сформированный набор данных был проанализирован на предмет количества повторений ошибок в словах из выборки, составлен словарь распределения наиболее часто встречающихся опечаток с весовыми коэффициентами для каждой буквы алфавита. Частотность, при этом, являлась основным критерием в процессе назначения весовых коэффициентов.

Результирующая выборка была дополнена данными об операторе (пол, возраст, род деятельности), производящем набор текста с целью дальнейшей кластеризации данных. Результирующий набор данных с распределением весовых коэффициентов представлен в табл. 3.

Таблица 3

Результирующий набор данных

Заменяемый символ	Распределение символов (символ: весовой коэффициент)							
А	О:32	В:45	У:19	И:16	Я:9	М:32	Е:47	К:19
Б	П:44	Ь:12	Л:15	Ю:72	Т:5	Д:42	О:6	М:2
В	Ф:26	А:32	И:4	Ы:23	К:16	У:19	Ц:5	Ч:12
Г	Х:34	Ш:17	Н:31	О:19	Р:12	Т:3	Щ:2	Р:11
Д	Т:37	Л:36	Ж:21	Щ:12	Ш:7	З:11	Б:8	Л:9
Е	Ё:93	И:53	Э:9	Н:34	К:18	П:23	У:9	Ч:7

Получение данных в случае использования мобильных устройств с сенсорной клавиатурой имеет свои специфические особенности. С помощью разработанных программных средств на языке Python из открытых источников было произведено извлечение текстов, написанных пользователями с применением клавиатуры мобильного устройства в качестве источника ввода.

Затем с использованием словаря русского языка и реализации алгоритма нечеткого сравнения строк были найдены слова, написанные с ошибками, и подобраны наиболее подходящие замены из словаря. В случае если слово имеет 2 и более варианта словарного написания, оно исключалось из выборки. Пороговое значение максимум в 2 перестановки не позволило сильно отличающимся словам попасть в итоговую выборку. Таким образом, было проанализировано 10000 коротких текстов и выявлено 14769 слов, написанных с ошибками.

Составлен словарь распределения наиболее часто встречающихся опечаток с весовыми коэффициентами для каждой буквы алфавита. Основным критерием в процессе назначения весовых коэффициентов тоже являлась частотность.

6. Объединение результатов

Для выбора определенного варианта из двух или более слов при нечетком сравнении строк, необходимо применить к каждой паре слов вышеупомянутые алгоритмы и, получив значения в виде коэффициентов уверенности, выбрать наибольший из коэффициентов.

Для оценки качества рассматриваемого алгоритма был сформирован набор тестовых данных, включающий в себя варианты слов с опечатками и множественным выбором при нечетком сравнении строк. Для каждого слова выбран определенный вариант, соответствующий правильному написанию. Объем набора данных составлял 1000 слов. Так как алгоритм является линейным, для оценки качества работы алгоритма использовалась метрика точности, значение которой составило 84%.

При анализе результатов работы можно сделать вывод о корректном выборе варианта слов в 84 случаях из 100. Следует обратить внимание на то, что в используемых данных было исключено использование сленговых выражений и ненормативной лексики. Алгоритм совершает ложноположительное срабатывание в результате отсутствия контекста, ориентируясь на показатель частотности замены символов. Пример вывода представлен на рисунке.

```
удевиться -> удивиться
пришол -> пришёл
убаалцсш -> убралась
зделать -> сделать
объясныть -> объяснять
моыатп -> мотать
сувать -> совать
токлй -> такой
ошыпка -> ошибка
руском -> пуском
минусать -> миновать
плейя -> племя
бутер -> буфет
донат -> донат
торелка -> горелка
зделать -> сделать
класный -> красный
издеватся -> издеваться
```

Рисунок. Пример вывода алгоритма нечеткого сравнения строк

Заключение

Проведен вычислительный эксперимент, подтверждающий жизнеспособность алгоритма выбора варианта при нечетком сравнении строк. Для оценки качества работы алгоритма приведена метрика точности. Планируется применение в корпоративных продуктах с целью прогнозирования развития бизнес-процесса.

Литература

2. Николаев И. С. Прикладная и компьютерная лингвистика / И. С. Николаев, О. В. Митренина, Т. М. Ландо ; Москва: URSS, 2016. – 320 с.

3. Ингерсолл Г. С. Обработка неструктурированных текстов / Г. С. Ингерсолл, Т. С. Мортон, Э. Л. Фэррис ; Москва: ДМК Пресс, 2015. – 414 с.

4. Jaro M. Advances in record linkage methodology as applied to the 1985 census of Tampa Florida // Journal of the American Statistical Association. – 1989. – Vol. 84, No 406. – P. 414–420.

5. Winkler W. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage // American Statistical Association. – 1990. – P. 354–359.

6. Winkler W. Overview of Record Linkage and Current Research Directions // Research Report Series. – 2006.

7. Воронина И. Е. Компьютерное моделирование лингвистических объектов: мо-нография / И. Е. Воронина ; Воронеж : Издательско-полиграфический центр Воронежского государственного университета, 2007. – 177 с.