НЕЧЕТКОЕ СРАВНЕНИЕ СТРОК НА ОСНОВЕ АДАПТАЦИИ ФОНЕТИЧЕСКИХ АЛГОРИТМОВ К РУССКОМУ ЯЗЫКУ

П. В. Белых, В. В. Парамонов, А. О. Шигаров
Институт динамики систем и теории управления имени В. М. Матросова Сибирского отделения
Российской академии наук
Иркутск, Российская Федерация
E-mail: {slv, shigarov, polina}@icc.ru

В работе рассматриваются вопросы адаптации фонетических алгоритмов нечеткого сравнения строк. Использование преобразований основанных на фонетических правилах русского языка позволяет более полно обеспечить сопоставление слов на основе их фонетического сходства. Развитие методов эффективного фонетического анализа в совокупности с использованием алгоритмов нечеткого сравнения строк, позволяет повысить качество сопоставления строк, например, для интеграции разнородной текстовой информации.

Введение

Информация, полученная из различных источников, предназначенная для интеграции в базе данных и последующего анализа, требует очистки. Очистка данных это процесс выявления и исправления ошибок и несоответствий с целью улучшения качества данных [1]. Процесс очистки данных включает в себя множество аспектов, таких как выявление орфографических ошибок, пропуска данных, наличия фиктивных, закодированных, составных значений, логических несоответствий. Использование фонетических алгоритмов для сравнения слов является одной из составляющих процесса очистки данных.

Фонетическое кодирование слов позволяет повысить качество их сравнения при различном написании. Основное назначение фонетических алгоритмов – определение оценки схожести слов на основе их фонетического сходства. Все существующие на сегодняшний день фонетические алгоритмы используют кодирование слов в зависимости от особенностей произношения слова, а не от орфографических правил его написания. В большинстве случаев, фонетические алгоритмы используются для сопоставления фамилий. Нам представляется, что достаточно эффективно использовать фонетические алгоритмы для сопоставления пользовательских строк с эталонными значениями различных классификаторов.

I. Применение фонетических алгоритмов

Применение только фонетических алгоритмов для решения задач очистки текста достаточно малоэффективно. Так как полученный код не дает возможности оценить не связанную с фонетическими правилами схожесть двух слов. Однако, при комплексном применении фонетических алгоритмов в совокупности с алгоритмами нечеткого сравнения строк, возможно повысить качество выявления и автоматического устранения орфографических ошибок в исходных (сырых) данных. Под комплексностью понимается использование фонетических алгоритмов для более качественной идентификации слова, если методы нечеткого сравнения, такие как метрика (расстояние) Левенштейна, строк показали результат в рамках заданного диапазона [2].

В настоящее время широко используются такие фонетические алгоритмы, как Soundex и его модификации, NYSIIS, Double Metaphone, Caverphone. Практически все фонетические алгоритмы, ориентированы на использование фонетических правил английского языка. Реализованы некоторые модификации алгоритмов [3], например, для французского, испанского языков. Существуют и другие адаптации фонетических алгоритмов для языков, основой которых не является латиница. Как правило, в этом случае используется транслитерация, а в качестве алгоритма используются вариации Soundex, Mrtaphone [3, 4]. Однако, транслитерация в ряде случаев не позволяет учесть особенности фонетики искомого языка.

II. Нечеткое фонетическое сравнение строк на русском языке

В предлагаемом подходе фонетические алгоритмы применяются для нечеткого сравнения строк на русском языке. Для кодирования строки используется простые числа. Результат кодировки -– сумма простых чисел. За основу кодирования взят алгоритм [5], в который внесены некоторые изменения. Так, например, нет условия уникальности буквы в слове. Исключение составляют сдвоенные буквы («нн», «оо» и т.п.), так как сдвоенность не всегда можно определить на слух.

При этом перед кодированием проводится процедура трансформации строки исходя из её фонетических особенностей, что позволяет получить код, максимально подходящий для созвучных строк. Все буквы слова переводятся в нижний регистр; удаляются все символы, не принад-

лежащие алфавиту; удаляются буквы «ъ», «ь»; проводится замена букв, дающих созвучные звуки, например: «а», «ё», «о» → а; «б», «п» → «п». Также проводится обобщение парных букв, что позволяет кодировать слово в зависимости от его звучания. Таким образом, формируется написание слов в соответствии с их звучанием. В ряде случаев звучание зависит от ударения. В настоящей реализации алгоритма данная особенность не учитывается.

Некоторые последовательности букв дают отличные звуки, например: «тс», «тц» → «ц»; «лнц», «ндц» → «нц», в связи с чем проводится модификация строки. Подобного рода модификация позволяет учесть особенности произношения слова при наличии определенного сочетания букв.

В целом, суть алгоритма состоит в модификации слова, обработке определённого количества символов и суммировании их кодов. Это дает возможность представлять результат кодировки вне зависимости от того какой символ последовательности обрабатывается, что позволяет избежать опечаток, связанных с переменой рядом стоящих местами букв.

III. Экспериментальное сравнение с другими алгоритмами

Для апробации был использовано множество из 300 различных слов. Количество слов на выходе – порядка 10000. В слова были внесены фонетические ошибки в которых учитывались учитывались комбинаторные и позиционные фонетические изменения, а также такие фонетические процессы как диэреза, фузия и метотеза, приводящие к появлению ошибок. В комбинаторных изменениях мы рассматривали ассимиляцию. Явление ассимиляции заключается в уподоблении звуков, т.е. взаимодействующие произносимые звуки становятся ближе полностью или частично (ножка [шк], отдать [дд], сдоба [зд], косьба [зьб]). При позиционных изменениях звучание звука зависит от его положения в слове (в абсолютном начале или конце слова) и отношением к ударению. Позиционное оглушение и озвончение согласных: звонкие парные оглушаются в конце слова и перед глухими согласными (мозг [ск], параход [т]). Глухие согласные озвончаются, в случае их расположения перед звонкими (за исключением непарных звонких согласных и «в») (сдать [здать]).

При диерезе - один звук выкидывается и образуется другой звук (сердце [с'эрцъ], солнце [сонцэ]). При фузии происходит слияние согласных звуков: (жарится моется — жарит(ц)а, мыться — мы(ц)а). При метатезе происходит взаимная перестановка звуков или слогов в словах (тарелка – талерка, молоток – мотолок).

Результаты применения адаптированного алгоритма сравнивались с результатами алгоритмов Soundex, Metaphone, Caverphone, Daitch–Mokotoff Soundex. Транслитерация выполнялась в соответствии с ГОСТ Р 52535.1–2006 [6]. В результате проведения теста предлагаемый алгоритм обработал 100% модификаций слов. В результате применения алгоритмов Soundex, Metaphone над множеством тестовых слов было корректно идентифицировано 90% , в случае же алгоритмов Caverphone, Daitch – Mokotoff Soundex — порядка 95%.

IV. Выводы

В предлагаемой адаптации, в отличие от большинства используемых фонетических алгоритмов, не строится транслитерация, что позволяет учитывать особенности произношения слов русского языка. При наполнении базы фонетических преобразований алгоритм может быть распространен языки восточнославянской группы.

Фонетическое сопоставление слов может быть использовано для эффективного сопоставления пользовательского текста с информации, содержащейся в различных классификаторах. В совокупности с применением методов нечеткого сравнения строк, данный подход может использоваться в алгоритмах очистки данных и представляет интерес для создания интегрированной информационной среды.

Исследование выполнено при финансовой поддержке РФФИ в рамках научных проектов №15–37–20042, 15–47–04348.

1. Li Zhao, Sung Sam Yuan, Sun Peng, Ling Tok Wang A New Efficient Data Cleansing Method // Database and Expert Systems Applications. 13th International Conference, DEXA 2002 Aixen – Provence, France, September 2 – 6, 2002 Proceedings. 2002
2. Рубцов Д. Н. Барахнин В. Б. О возможности борьбы с дубликатами при запросах к разнородным библиографическим источникам / Труды Одиннадцатой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2009). Петрозаводск, 17-21 октября 2009 г. С.293–298.
3. Alotaibi Yousef, Meftah Ali Review of distinctive phonetic features and the Arabic share in related modern research // Turkish Journal of Electrical Engineering And Computer Sciences 2013, Vol. 21, Issue 5, pp. 1426–1439.
4. Howida A. Shedeed A New Intelligent Methodology for Computer based Assessment of Short Answer Question based on a new Enhanced Soundex phonetic Algorithm for Arabic Language // International Journal of Computer Applications. Vol. 34, No. 10, November 2011. pp. 40–47.
5. Ставровиецкий В. В., Гапанюк Ю. Е., Галкин В. А. Алгоритм нечеткого фонетического поиска на основе простых чисел. [Электронный ресурс]: Молодежный научно-технический вестник №07, июль 2012 http://sntbul.bmstu.ru/file/505518.html?_s=1 – Дата доступа : 08.08.2015.
6. ГОСТ Р 52535.1–2006. Карты идентификационные. Машиносчитываемые дорожные документы. Часть 1. Машиносчитываемые паспорта. Национальный стандарт Российской Федерации. М.: Стандартинформ. – 2006. 18 с.