электронное научно-техническое издание
НАУКА и ОБРАЗОВАНИЕ
Эл № ФС 77 - 30569. Государственная регистрация №0421100025. ISSN 1994-0408

Распознавание текстового изображения с учетом морфологии слова

77-30569/350020

04, апрель 2012

Рудаков И. В., Романов А. С.

УДК 004.93

МГТУ им. Н.Э. Баумана
irudakov@yandex.ru

Введение.

Задача распознавания текстовой информации при переводе печатного и рукописного текста в электронный вид является одной из важнейших составляющих проектов, имеющих целью автоматизацию документооборота. Сложность считывания печатных документов заключается в необходимости обеспечить высокую надежность распознавания (более 98-99 %) даже при плохом качестве печати и оцифровки исходного текста.

В настоящее время, благодаря использованию компьютерных технологий, были развиты новые методы обработки изображений и распознавания образов [3], благодаря чему стало возможным создание таких систем распознавания печатного текста, которые удовлетворяли бы основным требованиям систем автоматизации документооборота. Однако перед приложениями по-прежнему ставятся задачи увеличения быстродействия и качества распознавания, минимизации затрачиваемой памяти, что требует дополнительных исследований в данной области.

Многие современные системы не учитывают структуру языка, на котором написан документ, а эти данные необходимы для последующей обработки ошибок. Для успешного процесса коррекции важны эффективные алгоритмы диагностики грамматических ошибок. В общем случае все сводится к определению принадлежности последовательности символов к данному естественному языку. Исправление опечаток определенных классов, в том числе однобуквенных, является практически важной задачей. Алгоритмы исправления ошибок в русских словах должны учитывать особенности русского языка как высоко флективного.

Предлагается метод распознавания текстового изображения с учетом морфологического анализа слова и разработка программы, реализующей этот метод.

Выбор алгоритма исправления ошибок в слове.

В процессе разработки было рассмотрено три наиболее используемых алгоритма [3] исправления ошибок в слове: расстояние Левенштейна, метод полных обратных преобразований и поиск максимальной подпоследовательности. Расстояние Левенштейна и метод поиска максимальной подпоследовательности дают очень хорошие результаты при коррекции, однако имеют сложность зависимости от словаря больше линейной. [3] Потому в работе был использован метод полных обратных преобразований.

Метод полных обратных преобразований подразумевает, что в слове содержится не более одной ошибки, следовательно, для ее исправления (применительно к задаче распознавания текстовых изображений) необходимо изменить каждый символ в слове и полученную словоформу проверить на наличие в словаре. Если словоформа присутствует в словаре, то она заносится в список корректных кандидатов.

Таким образом, обеспечивается высокая вероятность коррекции ошибок, если корректное слово имеется в словаре. Особенностью алгоритма является то, что обрабатываемые токены никак не оцениваются, а потому невозможно выбрать наиболее подходящий вариант для исправления ошибки, т.е. требуется вмешательство оператора.

Для наиболее корректной диагностики грамматических ошибок необходимо более корректно учитывать структуру и особенности языка. Таковым является морфологический анализ. [4]

Морфологический анализ слова.

Учет морфемной структуры слов позволяет компактно представлять совокупность словоформ, группируя их в словообразовательные гнезда. [5] Такое представление реализуется в виде словаря морфем, содержащего три части – корневую, префиксальную и суффиксальную.

В корневой части словаря корни расположены в лексикографическом порядке. В словарной статье под каждым заглавным корнем приводятся однокоренные слова, расчлененные на морфемы. Они расположены следующим образом. Во главе гнезда ставится корневое слово, причем на первом месте обычно располагается существительные с нулевым или выраженным окончанием, на втором –

неизменяемые части речи типа наречий, междометий. За корневыми словами идут беспрефиксные слова, имеющие суффиксы (лексикографически упорядоченные), затем следуют по алфавиту префиксальные слова первого префикса в слове.

Если при одной и той же основе может быть несколько производных, то непосредственно после основы слова, разделенной на морфемы, указываются под различными верхними индексами окончания, расположенные в алфавитном порядке, которые, присоединяясь к основе, образуют различные слова.

В префиксальной части словаря под каждым заглавным префиксом дается в алфавитном порядке перечень всех аффиксальных окружений корня, в которых встречается данный префикс; рядом перечисляются все корни, в которых встречается данный префикс; рядом перечисляются все корни, употребляющиеся в соответствующем окружении. Слова, начинающиеся прямо с корня приводятся в начале. [3]

Структура используемого словаря.

Выбранная структура эквивалентна структуре словаря Кузнецова А.И., Ефремова Т.Ф. и представляет собой текстовый файл в особом формате. Главная секция представляет набор структур, содержащая префикс и постфикс. Ещё одна секция представляет набор корней с указателями соответствующую структуру. Последняя секция представляет собой набор иноязычных приставок. Таким образом, достигается приемлемый процент сжатия словаря по сравнению с простым перечислением словоформ.

В начале реализации метода коррекции грамматических ошибок на основе морфологического анализа была использована реляционная база данных (на основе продукта MySQL) для хранения словаря. Однако проведенные тесты показали, что более 90% времени процессорного времени тратиться на разбор SQL-выражений, что побудило отказаться от использования реляционных баз данных. В результате была разработана структура для хранения словаря на основе деревьев.

Структура содержит три дерева. Первое дерево описывает все префиксы и иноязычные приставки. В конце каждого префикса имеется указатель на дерево корней, связанных с ним. Каждый корень имеет указатель на постфикс, принадлежащий одному классу с корнем и префиксом и связанный с этим корнем.

Создание специализированного хранилища для словаря позволило увеличить производительность и уменьшить сложность выборки слова для работы алгоритма коррекции грамматических ошибок с помощью морфологического анализа.

Зависимость точности исправления от положения ошибки в слове.

В ходе исследования влияния положения ошибок в слове на точность их исправления было определенно, что ошибки, допущенные во всех частях слова, кроме окончания, исправляются автоматически (рис. 1). Ошибки в окончаниях требуют помимо морфологического анализа еще и контекстный анализ, т.к. требуется выявить связь между словами. Для исследования был введен текст, содержащий ошибки во всех частях слова.

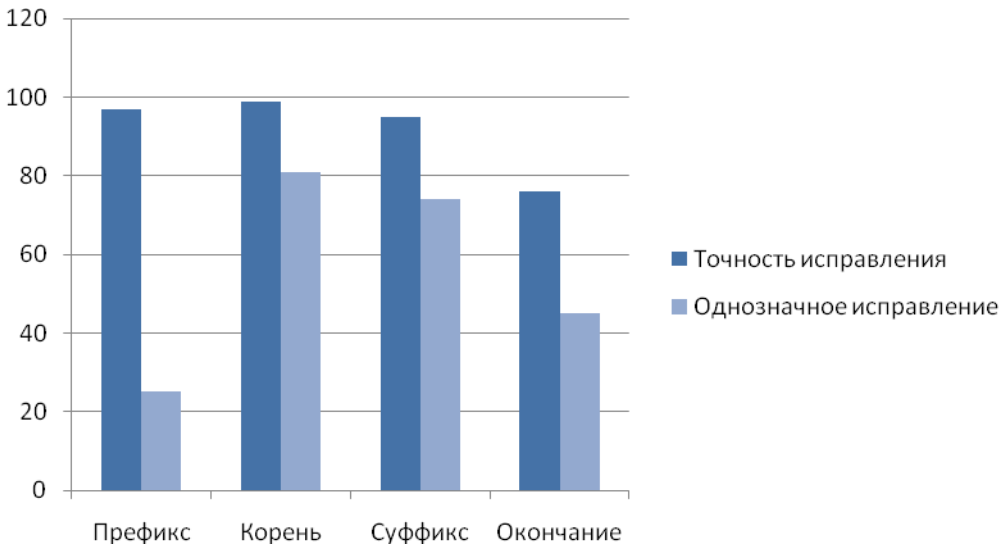

Рис. 1. Зависимость точности исправления ошибок от их положения в слове

Большое количество слов, полученных в результате исправления ошибок, не являются словоформами исследуемого слова и имеют лишь сходную с ним структуру. Это обуславливает невысокий процент однозначных исправлений. Тем не менее, автоматически исключить такие ошибки нельзя, т.к. не учитывается контекст.

В силу свойств разработанного алгоритма, было установлено, что если ошибка находится в префиксе, то время на ее исправление пропорционально длине слова на размер используемого словаря, что дает значительные временные затраты, так как для каждого рассматриваемого префикса генерируется полное обратное преобразование остальной части слова, количество однозначных исправлений ошибок в префиксе невелико. Быстрее всего ошибки исправляются в корне, т.к. на первом этапе алгоритма ищутся ошибки в основе, и если были найдены возможные варианты исправления, то поиск заканчивается.

Выводы.

В результате был разработан программный комплекс, позволяющий распознавать текстовые изображения и исправлять ошибки, полученные в процессе распознавания. Для увеличения скорости работы алгоритма исправления ошибок был разработан и реализован алгоритм разбиения исследуемого слова на морфемы, а так же разработана своя структура быстрого доступа к словарю. Для самого распознавания изображения была использована нейронная сеть прямого распространения и реализован алгоритм обучения, основанный на обратном распространении ошибки.

В дальнейшем для более точного исправления ошибок планируется реализовать помимо морфологического анализа слова еще и контекстный анализ, что значительно улучшит точность исправления ошибок, т.к. будут учитываться связи между словами, что позволит давать качественные оценки словам-кандидатам на основе их связи с контекстом всей фразы.

Литература

1. Арлазаров В.Л., Астахов А.Д., Троянкер В.В., Котович Н.В. «Адаптивное распознавание символов». Изд.: Интеллектуальные технологии ввода и обработки информации, Москва, 2001 г. – 580 стр.

2. Фисенко В.Т., Фисенко Т.Ю. «Компьютерная обработка и распознавание изображений». Изд.: СПбГУ ИТМО, СПб, 2008 г. – 756 стр.

3. Гниловская Л.П. Гниловская Н.Ф. «Автоматическая коррекция орфографических ошибок». Изд.: Мир, Москва, 1984 г. – 278 стр.

4. Бутакова Л.О. Опыт классификации ошибок, свойственной письменной речи. 1998. Internet: http://www.omsu.omskreg.ru/vestnik/articles/y1998-i2/a072/article.html

5. Кузнецова А.И., Ефремова Т.Ф. «Словарь морфем русского языка». Изд.: Русский язык, Москва, 1986 г. – 1134 стр.

electronic scientific and technical periodical
SCIENCE and EDUCATION
EL № FS 77 - 30569. №0421100025. ISSN 1994-0408

Recognition of text image subject to word morphology

77-30569/350020

04, April 2012

Rudakov I.V., Romanov A.S.

Bauman Moscow State Technical University
irudakov@yandex.ru

Algorithms of auto-correction in documents were analyzed. The authors proposed a method of correcting grammar mistakes by means of a morphological analysis. The authors present a bundled software allowing to recognize text images by means of neural network and to correct mistakes with the usage of the proposed method. Creation of a special storage for the dictionary allowed to increase productivity and decrease the complexity of word selection for the operation of the bundled software. The dependence of correction accuracy on the mistake location in the word was identified.

Publications with keywords: text recognition, word-lore analysis, Levenshtein distance, method of general inverse transformations, search of maximal sequence
Publications with words: text recognition, word-lore analysis, Levenshtein distance, method of general inverse transformations, search of maximal sequence

References

1. Arlazarov V.L., Astakhov A.D., Troianker V.V., Kotovich N.V. *Adaptivnoe raspoznavanie simvolov* [Adaptive character recognition]. Moscow, Intellektual'nye tekhnologii vvoda i obrabotki informatsii, 2001. 580 p.

2. Fisenko V.T., Fisenko T.Iu. *Komp'iuternaia obrabotka i raspoznavanie izobrazhenii* [Computer processing and recognition of images]. SPb, SPbGU ITMO Publ., 2008. 756 p.

3. Gnilovskaia L.P., Gnilovskaia N.F. *Avtomaticheskaia korrektsiia orfograficheskikh oshibok* [Automatic correction of spelling errors]. Moscow, Mir, 1984. 278 p.

4. Butakova L.O. *Opyt klassifikatsii oshibok, svoistvennoi pis'mennoi rechi* [Experience of classification of errors inherent in written speech]. 1998. Available at: http://www.omsu.omskreg.ru/vestnik/articles/y1998-i2/a072/article.html.

5. Kuznetsova A.I., Efremova T.F. *Slovar' morfem russkogo iazyka* [Dictionary of morphemes of the Russian language]. Moscow, Russkii iazyk, 1986. 1134p.