11. *Канеман Д.* Думай медленно. Решай быстро. – М. : АСТ, 2014. – 315 с.
12. *Кропачев Л. А.* Указ. соч.
13. Там же. С. 8.
14. *Моченов С. В.*, *Шаронов М. А.*, *Ахметгалеев Р. Р.* Указ. соч.
15. *Кропачев Л. А.* Указ. соч.
16. *Моченов С. В.*, *Шаронов М. А.*, *Ахметгалеев Р. Р.* Указ. соч.
17. *Кропачев Л. А.* Указ. соч.
18. *Ревенков А. В.*, *Резчикова Е. В.* Указ. соч.
19. Там же.
20. *Кропачев Л. А.* Указ. соч.
21. *Распопов И. П.* Строение простого предложения. – М., 1970. – 191 с.
22. *Ковтунова И. И.* Порядок слов и актуальное членение. – М., 1976. – 239 с.
23. *Моченов С. В.*, *Шаронов М. А.*, *Ахметгалеев Р. Р.* Указ. соч.

Получено 05.10.2015

УДК 004.93

И. В. Сабуров, магистрант, ИжГТУ имени М. Т. Калашникова
А. В. Кучуганов, кандидат технических наук, ИжГТУ имени М. Т. Калашникова
М. Н. Мокроусов, кандидат технических наук, ИжГТУ имени М. Т. Калашникова

ПРИМЕНЕНИЕ СЛОВАРЕЙ В ЗАДАЧЕ РАСПОЗНАВАНИЯ РУКОПИСНЫХ ТЕКСТОВ *

В настоящее время проблема offline-распознавания рукописного текста не решена. Программы для распознавания текста применяются в компьютерной лингвистике, а также в задачах, объектом которых является рукописный текст. Предлагается подход к повышению качества и надежности распознавания рукописного текста за счет поиска подходящих слов в морфологическом словаре Зализняка.

Рассмотрим несколько словарей русского языка, которые могут быть использованы в программе распознавания рукописного текста.

Словообразовательные (деривационные) словари – словари, показывающие членение слов на составляющие их морфемы (минимальные значимые части слова), словообразовательную структуру слова, а также совокупность слов с данной морфемой – корневой или аффиксальной [1]. Слова в словообразовательных словарях приводятся с членением на морфемы и с ударением.

Существует четыре основных типа морфемных словообразовательных словарей: словари-корнесловы (единицами таких словарей являются корневые морфемы, в алфавитном порядке приводятся слова без указания на словообразовательные отношения однокоренных слов); словари морфемной членимости слов (задача таких словарей – показать не только морфемный состав каждого слова, но и раскрыть его словообразовательную структуру); толковые словари аффиксальных морфем (такие словари раскрывают значение аффиксов и особенности их функционирования); частотные словообразовательные словари (морфемы расположены по их убывающей частотности).

Морфологический словарь или словарь словоформ русского языка содержит полную акцентуированную парадигму слов русского языка с их полным морфологическим описанием. С помощью специальной системы условных обозначений словарь отражает современное словоизменение, то есть склонение существительных, прилагательных, местоимений, числительных и спряжение глаголов. Задача данного словаря – раскрыть морфологический потенциал слова.

Для работы с программой наиболее удачным и подходящим словарем является морфологический словарь, так как в отличие от морфемного словаря в нем приводятся полные и нечленимые формы слов, а также полная парадигма слова.

В программе использовался морфологический словарь Андрея Анатольевича Зализняка. Данный словарь содержит около 100 тысяч лексем, каждая из которых включает в себя полную парадигму словоформ и информацию, позволяющую построить любую грамматически правильную форму любого из этих слов.

Словарь академика Зализняка – основополагающий труд по морфологии, где впервые был предложен системный подход к описанию грамматических парадигм, включающих не только изменение буквенного состава слов, но и ударения [2]. Электронная версия этого словаря легла в основу большинства современных компьютерных программ, работающих с русской морфологией: системы проверки орфографии, машинного перевода, автоматического реферирования и т. д.

Рассмотрим укрупненный алгоритм использования морфологического словаря при распознавании рукописного текста.

После этапа графического распознавания каждая буква содержит несколько потенциальных вариантов, которые достоверны с определенной вероятностью. Данная вероятность отражает процент совпадения

* Работа выполнена при частичной поддержке Госзадания МОиН РФ (проект № 625) и гранта РФФИ № 15-07-08077 А.

графа эталона с графом буквы, построенным на этапе предобработки графического распознавания.

Далее для каждого слова выполняется побуквенный поиск в словаре. Существует несколько известных алгоритмов нечеткого поиска по словарю: расстояние Левенштейна, расстояние Дамерау – Левенштейна, алгоритм расширения выборки, метод N-грамм, хеширование по сигнатуре, BK-деревья.

Необходимо отметить, что почерк вовсе не представляет собой что-либо застывшее, раз и навсегда определившееся, а напротив, может претерпевать изменения и иногда довольно существенные. Изменения окружающих условий, настроения, внезапные эмоциональные реакции (состояния возбуждения, подавленности и т. п.) зачастую существенно влияют на почерк. В задаче распознавания рукописных текстов выделяют следующие крупные проблемы: бесконечное количество разновидностей почерка, раздельное написание некоторых элементов слова или наличие декоративных элементов; сложность выявления отдельных символов в слитном рукописном слове; зависимость написания символа от его положения в слове.

По этой причине после этапа распознавания в результирующих последовательностях возможны пропуски символов, появление новых символов. Для решения проблемы мы применили простой рекурсивный алгоритм побуквенного поиска, в котором на каждой итерации используется результат выборки по предыдущим последовательностям вариантов, а в случае пустой выборки происходит возврат к предыдущим результатам поиска.

Суть алгоритма поиска: берется первый символ слова, и выбираются из словаря те словоформы, которые начинаются с любых вариантов этого символа. Затем в данной выборке ищутся те словоформы, в которых вторые буквы совпадают с вариантами второго символа распознанного слова. Далее сравниваются третья буква в словоформах выборки и варианты третьего символа, и так далее до тех пор, пока в выборке не останется одно слово или не будут рассмотрены все символы. Если в выборке остается одно слово, то для каждого символа останется только один вариант. Если выборка пустая, то осуществляется возврат к предыдущему символу и результату, к вариантам символа добавляются все возможные буквы алфавита, и поиск повторяется, пока не будет выполнено условие остановки.

В результате такого побуквенного поиска по словарю сокращаются варианты распознавания символа, удаляются те варианты, которых нет в итоговой выборке.

Ниже на рисунках представлены результаты распознавания рукописной фразы в программе распознавания рукописного текста Reco2 [3]. В программе можно задать такие настройки, как минимальный процент совпадений, минимальное количество кандидатов (вариантов) распознавания, глубина поиска соседних символов, погрешность ориентации.

На рис. 1 показан пример распознавания текста до применения грамматического словаря.

Как видно по рис. 1, в результат распознавания попали слова, которые написаны с ошибкой.

В результате уточнения по словарю, в котором побуквенно провели поиск каждого распознанного слова, некорректные варианты распознавания символов были отброшены и оставлены только те, которые образуют слово из словаря.

На рис. 2 показан результат распознавания с уточнением по морфологическому словарю. В левом среднем окне можно проследить за ходом поиска букв по словарю.

Таким образом, подключение базы данных Зализняка А. А., которая содержит список словоформ русского языка, позволило повысить качество и надежность распознавания рукописного текста за счет сокращения нерелевантных вариантов путем побуквенного поиска подходящих слов.

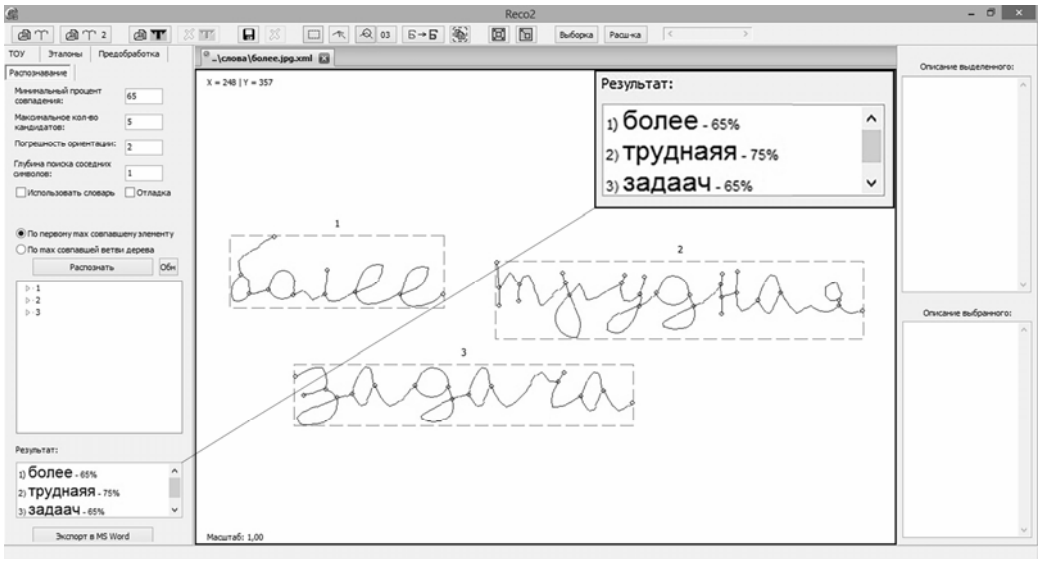

Рис. 1. Пример распознавания фразы без обращения к словарю

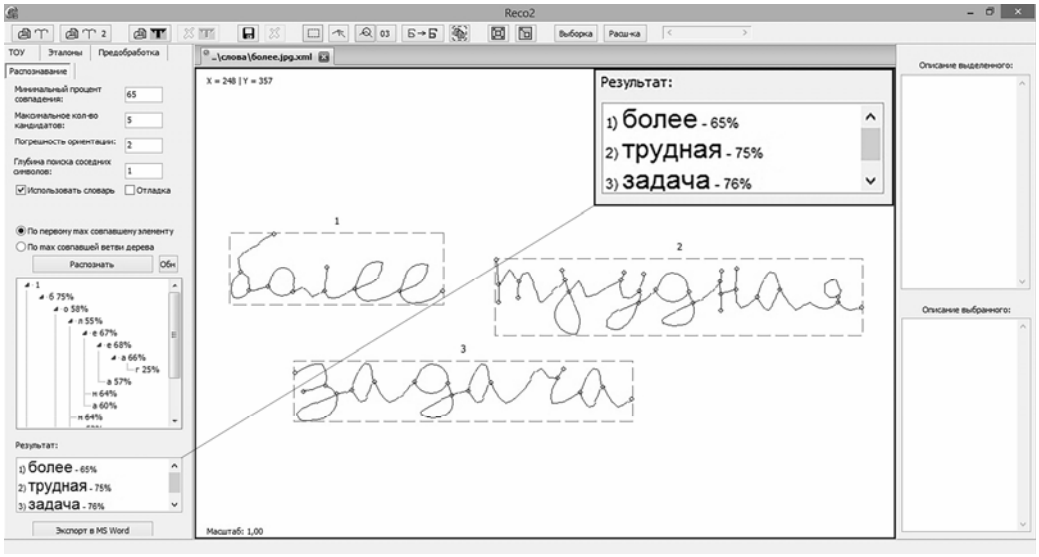

Рис. 2. Пример распознавания фразы с обращением к словарю

Библиографические ссылки

1. Какие бывают словари // Справочно-информационный портал «Русский язык». GRAMOTA. – URL: http://www.gramota.ru/

2. *Успенский В. А.* О русском языке, о дешифровке древних текстов, о «Слове» // Новый мир. – 2007. – № 8.

3. *Касимов Д. Р.*, *Кучуганов А. В.* RECO – программная система для распознавания старославянских текстов // Информационные технологии и письменное наследие : материалы междунар. науч. конф. (Уфа, 28–31 окт. 2010 г.) / отв. ред. В. А. Баранов. – Уфа ; Ижевск, 2010. – С. 144–148.

Получено 21.10.2015