Многопроходное распознавание смешанных печатных текстов на примере русско-английского распознавания

О. А. Славин

В статье описаны алгоритмы распознавания печатных многоязычных текстов, рассматриваемые на примере распознавания русско-английских документов. Рассмотрено применение как шрифтонезависимых, так и шрифтозависимых документов. Уделено внимание применению самообучения на основе кластеризации образов распознанных символов и использованию лингвистических механизмов.

Введение

Представляет интерес распознавание образов многоязычных текстов. В двуязычных текстах, коды которых принадлежат одной кодовой странице, включающей алфавиты обоих языков, например, английского и финского, отличия состоят в нескольких специфичных для финского языка буквах **ääöÄÖ**, возникающие конфликты между специфичными финскими символами и их английскими аналогами ничем не отличаются от таких же конфликтов в рамках одного финского языка. Алфавит финского языка, рассматриваемый как расширение латиницы, позволяет распознавать не только финские, но английские слова.

Не такова ситуация для смешанного русско-английского текста. Деловые или научные русско-английские документы составлены на базе кириллицы и правил русского языка, но включают в себя отдельные английские слова. Алфавиты кириллицы и латиницы имеют часть символов, общих по начертаниям, и специфичные образы символов. Особенностью деловых русско-английских текстов является значительная доля не словарных английских слов (аббревиатур, терминов, названий торговых зна-

ков). Двуязычный русско-английский режим (далее под двуязычным мы рассматриваем только русско-английский режим) требует специальных алгоритмов для качественного распознавания печатных документов.

1. Шрифтонезависимое распознавание русско-английских текстов

Алгоритмы распознавания отдельных символов, как высокоточные нейронные сети [1], так и методы на основе эталонов [2] снижают качество распознавания при расширении алфавита из-за появления дополнительных родственных двуязычных символов. Например, нейронная сеть, описанная в работе [1], при распознавании последовательности из 1 000 000 прописных и строчных русских букв различных шрифтов уменьшает точность распознавания с 99,99 % (с алфавитом, содержащим только русские буквы) до 96 % (с алфавитом, содержащим латинские и русские буквы). Локализация образов символов с неизвестными границами (сегментация границ) также производится точнее на одном из алфавитов, нежели на объединенном алфавите.

Поэтому двухпроходное распознавание строк текста, на каждом из которых допустим только один алфавит, обеспечивает более точную сегментацию и распознавание символов. Возможен способ распознавания строк, состоящий в распознавания одной строки два раза с различными алфавитами, после чего производится объединение распознанных слов. Реальная доля английских слов в деловых русско-английских текстах составляет 5–10 %. Ввиду этого выгодно провести селекцию слов после первого русскоязычного прохода с целью нахождения части строки, которая может быть проинтерпретирована как английская, найденные кандидаты распознать по-английски как отдельную строку.

В русской и английских строках производится поиск и сопоставление слов, соответствующих одним и тем же последовательностям образов, и принимается решение о выборе одного из вариантов распознавания одного слова различными языками. Понятие *слова* определим как последовательность символов, отделенная от подобных себе последовательностей пробелами и иными разделителями (многочисленными знаками пунктуации и служебными символами, такими как **.,;:-“«»/{}()**), снабженными оценками надежности. Разделители с низкими оценками и знаки препинания могут быть возникать как по объективным причинам, но и как следствие ошибок распознавания. Последнее может потребовать проверки при сопоставлении двух разноязычных слов нескольких гипотез о границах слов. Можно сказать иначе, что при сопоставлении

границ двух слов из различных слов при сравнении результатов распознавания проходов с различными языками необходимо рассматривать варианты границ слов. Например, пара последовательно расположенных слов, связанных разделителем с низкой оценкой, может рассматриваться как одно слово.

Для выделения слов-кандидатов для перераспознавания и сравнения двух слов необходимо иметь оценки качества распознавания слова. Прежде, чем их сформулировать, обратимся к анализу смешанного алфавита, включающего в себя и кириллицу, и латиницу, с точки зрения неразличимых и родственных символов. В смешанном алфавите присутствует немалое число *неразличимых* (в шрифтнезависимом распознавании) символов, таких как **ETIOPAHMeyopalxcbm**. Группа *родственных* двуязычных символов также велика, к ней, a priori, относятся буквы **Yrnuv*Угпич**, некоторые символы могут быть надежно отсегментированы в другом алфавите иным способом, например, русский символ **Ы** равноценен паре английских символов **bI**. Перечень и конкурентоспособность родственных символов во многом зависят от свойств метода распознавания символа. Тем не менее, учет оценок родственных символов является ненадежным способом их отнесения к одному из алфавитов. Необходимо учесть неразделимость и конкурентоспособность некоторых символов с их аналогами в другом алфавите. Представляется разумной система характеристик качества распознавания слова в целом:

- минимальная и максимальные оценки отдельных символов;
- минимальная и максимальные оценки отдельных символов, за вычетом неразделимых символов;
- минимальная и максимальные оценки отдельных символов, за вычетом неразделимых и родственных символов;
- число нераспознанных символов;
- словарное подтверждение;
- оценки неоднородности слова: колебания прописных и строчных букв, букв и цифр;
- оценки неоднородности по отношению к соседним словам;
- оценка неопределенности границ слова.

Разумеется, помимо оценок отдельных символов могут потребоваться и оценки последовательностей, рассматриваемых как конкурент одного символа из другого языка. Сравнение слов двух языков и выделение кандидатов на перераспознавание другим языком осуществляется на основе перечисленных оценок f_i одного слова, объединенных в функцию сравне-

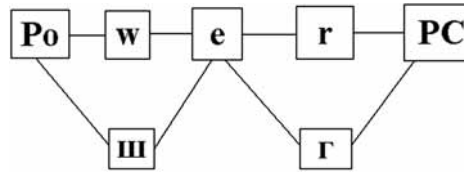

Рис. 1. Пример конфликтного сравнения двух слов

ния F (например, линейную). Веса функции сравнения оптимизируются на стендах двуязычных текстов, содержащих большое число конкурирующих слов.

При сравнении важно выделять случаи, когда отсутствуют дополнительные обстоятельства (словарь, колебания характеристик символов), позволяющие дать приоритет при отнесении слова к одному из языков, и решение можно принять только по оценкам символов, являющихся уникальными для одного из языков. Например, в паре слов ***РошегРС** и **PowerPC** символы **Po_e_PC** является неразличимой для омнифонтовых алгоритмов, а символы **_г_** и **_r_** — родственными, то есть ненадежно различаемыми. Возможность различения этих слов состоит в сравнении оценок символов **ш** и **w**, если же оценки обоих ненадежны (малы), то слова являются несравнимыми на данном этапе. Для последующего устранения конфликта необходимо запомнить оба слова, при их хранении в виде графа (см. рис. 1) сохраняются и результаты анализа символов слов с точки зрения их двуязычия. А именно, запомнятся типы пар символов **гr**, как родственных, и **шw**, как конфликтующих символов.

Для конкурирующих двуязычных пар слов, состоящих из одинакового числа образов, возможна более простая функция сравнения F_1, учитывающая значения оценок надежности соответствующих образов символов. Сравнение идет следующим путем. Для каждого символа из одного слова выбирается соответствующий символ из другого. Подсчитывается количество оценок надежности символов одного слова, превысивших оценки соответствующих им символов второго слова.

Функция F_1 обобщается для случая, когда пара слов состоит из различного числа символов, но различие незначительно. Соответствие устанавливается по границам образов распознанных символов. Функция F_2 призвана оценить пары слов, в которых некоторым широким символам одного слова может быть установлено соответствие нескольких образов другого слова.

2. Адаптивное распознавание русско-английских текстов

Алгоритмы адаптивного распознавания [3] могут работать не только с одним языком, она успешно работает и в случае, когда на странице одновременно присутствует текст на двух языках — русском и английском.

Самообучение в условиях русско-английского смешанного распознавания усложняется учетом ситуаций слов с неоднозначно найденным языком. Образы символов из слов, подобных слову на рис. 1 с неразрешенным конфликтом, не могут образовывать самостоятельных кластеров, трактуемых как надежные, а могут только пополнять кластеры образованные символами, язык которых выбран надежно. Построенная система эталонов должна быть проанализирована на предмет разрешения конфликтных ситуаций. Например, пары символов «rr» и «ш» могут быть разделены надежно сформированным эталоном. Вообще говоря, кластерное наложение способно в некоторых случаях различать и неразличимые омнифонтовыми методами образы при сборке надежно разделимых эталонов, подобно тому, как устраняется неразличимость символов в рамках одного языка.

В случае русско-английского варианта распознавания после первого прохода для каждого слова устанавливается признак уверенного определения языка.

Если язык слова определен, то кластерное распознавание идет обычным образом, но только по символам указанного языка. Кластеры иного языка в перераспознавании не участвуют.

Если по результатам распознавания на первом проходе для слова язык не определен (есть вариант распознавания по-русски и есть вариант по-английски), слово на втором проходе распознается дважды. Сначала перераспознается русское слово по русским кластерам, результаты первого и второго проходов русского перераспознавания смешиваются стандартным образом. Затем таким же образом перераспознается английский вариант (но теперь только по английским кластерам). Полученные варианты сравниваются по величинам оценок с помощью функций F, F_1, F_2.

Заключение

Рассмотренное решение проблем русско-английских текстов, использующее монотонные оценки, многокритериальное сравнение и дораспознавание на втором проходе, применимо и к другим парам языков, таким как русско-финский, или казахско-английский, выводимым из кириллицы и латиницы.

Литература

1. *Гавриков М. Б.*, *Мисюрев А. В.*, *Пестрякова Н. В.*, *Славин О. А.* Об одном методе распознавания символов, основанном на полиномиальной регрессии // Автоматика и телемеханика. 2006. № 3. С. 119–134.
2. *Болотин П. В.*, *Корольков Г. В.*, *Славин О. А.* Методы распознавания грубых объектов // Развитие безбумажных технологий в организациях: Труды ИСА РАН. М.: URSS, 1999. С. 331–355.
3. *Арлазаров В. Л.*, *Котович Н. В.*, *Славин О. А.* Адаптивное распознавание // Информационные технологии и вычислительные системы. 2002. № 4. С. 11–22.