Коррекция орфографических ошибок в русских текстах, полученных после обработки программами распознавания, с использованием словаря, индексированного хэшами.

Лысенков Александр Сергеевич,
Мифтахов Рустам Рамисович

студент, студент

Факультет ЭиА Казанского (Приволжского) Федерального Университета,

Набережные Челны, Россия

E-mail: lysenkovas@mail.ru, rust.mif@mail.ru

В настоящее время существует множество алгоритмов поиска и исправления орфографических ошибок в тексте. Все они так или иначе реализуют поиск в заранее подготовленном словаре тех слов, которые либо в точности соответствуют словам из исходного текста, либо могут быть преобразованы в последние минимальным количеством действий (заменой, вставкой, удалением и перестановкой букв). Цель данной работы — продемострировать такой алгоритм поиска ошибок, который учитывает специфику проблемы корректного распознавания текста на изображениях.

В отличии от ошибок пользователя, в ошибках распознающего ПО можно проследить простую закономерность: при сохранении верного количества и верного порядка символов, некоторые из них могут быть заменены внешне похожими. Для учёта этой особенности каждому слову в словаре ставится в соответствие хэш той же длины.

Определение 1. *Хэш (здесь и далее) — шеснадцатеричное число, в котором каждый разряд отражает принадлежность символа соответствующего слова на той же позиции некой группе внешне схожих символов (рис. 1.).*

Таким образом вместо полного перебора всего словаря для нахождения соответствия требуется рассмотреть лишь малую группу слов с таким же хэшем, как у искомого слова. В случае, если исходное изображение особенно некачественно, выборку слов из словаря можно расширить, предпологая, что каждый символ может относиться не только к своей, но и к «соседней» группе (рис. 1).

Также одна из особенностей распознанных текстов — «выпавшие» буквы. Иногда, из-за бликов или затёртостей, символ может

быть проигнорирован распознающим ПО. В таком случае слово оказывается разделено надвое пробелом. Рассматриваемый алгоритм можно использовать для решения этой проблемы, проведя поиск по всем хэшам, которые можно получить, объединив два ошибочных слова неизвестным символом.

При правильной организации индекса поиск по хэшу показывает очень хорошие результаты по времени работы [1,2]. В данном случае индекс организован в виде дерева, где каждый элемент — это хэш, который длиннее хэша родительского элемента на один разряд (рис. 2).

Иллюстрации

1	2	3	4	5	6	7	8	9	A	B	C	D
а	е	б	в	ъ	г	п	и	ж	у	ш	д	р
э	ё	о	з	ы	т	л	й	к	ц	щ		ф
	с	ю	я	ь		м	н	х	ч			

7	8	6	2	D	1	6	A	D	1
л	и	т	е	р	а	т	у	р	а

Рис. 1. Таблица, определяющая правило составления хэша, и пример соответствия.

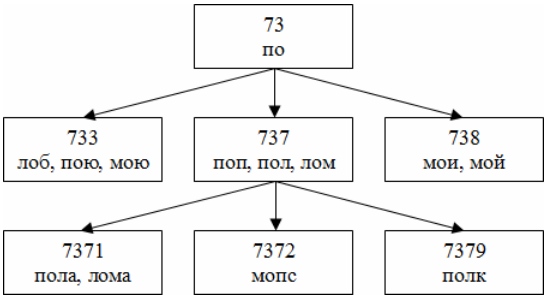

Рис. 2. Пример участка словаря.

Литература

1. Бойцов Л. М. Использование хеширования по сигнатуре для поиска по сходству. ВМиК, МГУ М. 2001, Прикладная математика и информатика. № 8 2001. С. 135-154.
2. Нгуен Н. Х. Обзор некоторых алгоритмов нестрогого сопоставления записей применительно к задаче исключения дублирования персональных данных. Молодой ученый. № 5 2013. С. 163-166.