Алгоритм подтверждения результатов распознавания с помощью словаря

О. А. Славин, И. М. Янишевский

Аннотация. В статье описывается способ подтверждения надежности распознанных символов с помощью корпуса словоформ. Приводится модель ошибок подтверждения и результаты вычислительных экспериментов для оценки вероятности ошибки словарного подтверждения. Обсуждаются вопросы комбинирования словарного подтверждения и результатов распознавания алгоритмов, формирующих монотонные оценки.

Ключевые слова: *Распознавание символов, образ символа, вероятности ошибки распознавания, словарь, словарное подтверждение.*

Введение

Современные алгоритмы распознавания текстов обладают высокой способностью распознавания текстовых образов, позволяя достичь высоких значений характеристик точности распознавания (доля правильно распознанных символов по отношению к общему объему символов) и монотонности оценок распознавания (доля ошибок с данной оценкой распознавания w по отношению к общему объему объектов, распознанных с этой оценкой w). Точные определения характеристик точности и монотонности оценок распознавания приведены в работе [2]. В статье рассматриваются алгоритмы распознавания образов символов (далее — алгоритмы распознавания), которые позволяют получить результат в форме (c, w), где c — код символа, а $w = w(c)$ — оценка распознавания (далее — оценки), наличие альтернатив распознавания и другой информации о распознавании не является обязательным. Для простоты будем считать оценки целочисленными и принадлежащими диапазону [0; 255], где 0 — минимальная из возможных оценок, а 255 — максимальная.

Рассмотрим задачу отбора надежно распознанных символов из множества распознанных образов символов. Естественное простое решение этой задачи состоит в отборе символов с оценками $w(c)$ с помощью набора заранее заданных порогов оценок $p_w(c)$ с помощью следующего правила: если $w(c) > p_w(c)$, то символ считается распознанным надежно.

Для алгоритмов распознавания с монотонными оценками приведенный способ отбора надежно распознанных символов приемлем, однако возникает проблема. Дело в том, что алгоритмы распознавания символов (для определенности — символов с заранее определенными границами), обладающие наивысшими значениями монотонности позволяют выделить лишь малую часть от объема правильно распознанных образов, т. е. $|\Im(p_w(c))| << |\Im(c)|$, где $\Im(p_w(c))$ — множество символов, каждый из которых отнесенных к надежно распознанным с помощью отбора по оценкам монотонности, а $\Im(c)$ — множество символов c, распознанных правильно. Например, алгоритм $\wp$ полиномиальной регрессии Пестряковой, описанный в [3], на тестовой последовательности из 1 000 000 символов печатного текста различного качества обеспечивает распознавания в полуинтервале оценок [240; 255] точность, равную 0,966. В то же самое время доля образов, распознанных алгоритмом $\wp$ и получивших оценки в полуинтервале [240; 255], составляет около 32 %.

Увеличение доли надежно распознанных образов возможно за счет комбинирования с другими механизмами. В настоящей работе рассматривается способ подтверждения надежности распознанных слов с помощью словаря (корпуса слов).

1. Модель словарного подтверждения

Исследуем возможности повышения надежности оценок распознавания символов с помощью словарей некоторым алгоритмом $\mathfrak{R}$, обладающим известными заранее характеристиками точности и монотонности оценок.

Результатом распознавания набора строк являются последовательности слов, причем известно разделение символов слов на прописные и строчные.

Известен *словарь* распознавания, т. е. набор словоформ определенного языка $V = V_1 \cup \ldots \cup Vm$, разби-

тый на секции $V_k = \{v = \alpha_1\alpha_{2\ldots}\alpha_k | \alpha_j \in A\}$ различной длины k, где $A = \{\alpha_1, \ldots, \alpha_n\}$ — алфавит распознавания, а также вероятность $p(\upsilon)$ встречаемости каждой из словоформ υ.

Для слова ω, полученного как результат распознавания образа слова алгоритмом $\mathfrak{R}$, рассмотрим процедуру *подтверждения словарем* слова ω, состоящую в поиске слова ω в массиве слов V. При этом будем рассматривать только распознавание слов из словаря V.

Для этого будем использовать распределение вероятностей ошибок распознавания алгоритма $\mathfrak{R}$ в форме $\{s_i, s_j, p_{ij}\}$, где $p_{ij} = p(s_i, s_j)$ — условная вероятность распознавания образа набора символов s_i как набора символов s_j, при этом оба набора символов состоят из букв алфавита A. Вообще говоря, длины s_i и s_j не обязаны совпадать. В частном случае s_i и s_j содержат ровно по одной букве.

Введем порог значимости p_0 вероятности распознавания с ошибкой: если $p_{ij} < p_0$, то полагаем $p_{ij} = 0$. Возможно, что существует несколько наборов символов s_i: $p_{ij} \neq 0$.

Зафиксируем слово $\upsilon = \alpha_1\alpha_{2\ldots}\alpha_k \in V$ и рассмотрим вероятные *прототипы* $\omega_z = \alpha'_1\alpha'_2\ldots\alpha'_m \in V$. Разобьем слово υ и прототип ω_z на части следующим образом:

$$\upsilon = s_i^{(\upsilon,1)} s_i^{(\upsilon,2)} \cdots s_i^{(\upsilon,q)},$$

$$\omega_z = s_j^{(\omega_z,1)} s_j^{(\omega_z,2)} \cdots s_j^{(\omega_z,q)},$$

где

$$s_i^{(\upsilon,1)} = \alpha_1 \cdots \alpha_{l^{(1)}}, \quad s_i^{(\upsilon,2)} = \alpha_{l^{(1)}+1} \cdots \alpha_{l^{(2)}}, \ldots,$$

$$s_i^{(\upsilon,q)} = \alpha_{l^{(q-1)}+1} \cdots \alpha_{l^{(q)}}, \; l^{(q)} = k, \quad s_j^{(\omega_z,1)} = \alpha'_1 \cdots \alpha'_{n^{(1)}},$$

$$s_j^{(w_z,2)} = \alpha'_{n^{(1)}+1} \cdots \alpha'_{n^{(2)}}, \ldots, \quad s_j^{(\omega_z,q)} = \alpha'_{n^{(q-1)}+1} \cdots \alpha'_{n^{(q)}},$$

$$n^{(q)} = m.$$

Процесс трансформации $\omega_z \to \upsilon$ заключается в замене $s_j^{(\omega_z,l)}$ на $s_i^{(\upsilon,l)}$, $l = \overline{1,q}$. Найдем вероятность трансформации слова ω_z в слово υ. Считая независимыми события, состоящие в замене $s_j^{(\omega_z,l)}$ на $s_i^{(\upsilon,l)}$, $l = \overline{1,q}$, получим необходимую формулу для вероятности:

$$p(\omega_z \to \upsilon) = \prod_{l=1}^{q} p(s_i^{(\upsilon,l)} \mid s_j^{(\omega_z,l)}),$$

где $p(s_i^{(\upsilon,l)} \mid s_j^{(\omega_z,l)})$ — вероятность распознавания образа набора символов $s_j^{(\omega_z,l)}$ как набора символов $s_i^{(\upsilon,l)}$. Вероятность $p(s_i^{(\upsilon,l)} \mid s_j^{(\omega_z,l)})$ отлична от нуля и известна заранее как один из элементов распределения $\{s_i, s_j, p_{ij}\}$.

Существование одного или нескольких прототипов является ошибкой подтверждения словарем.

Проверим факт наличия в словаре каждого из возможных прототипов ω_z. Если $\omega_z \in V_k$, то будем считать возможной ошибку распознавания словарного слова ω_z, причем с вероятностью $p(\omega_z \to \upsilon)$ это слово будет распознано как слово υ.

Для заданных заранее словаря и распределения вероятностей ошибок произведем эксперимент, состоящий в поиске возможных прототипов для всех словарных слов определенной длины и оценке вероятности ошибки подтверждения словарем. Найдем все различные слова ω, отображающиеся на другие словарные слова υ с вероятностью $p(\omega_z \to \upsilon)$.

Приведем формулу вычисления вероятности ошибки p_k в слове длины k:

$$p_k = \sum_{\upsilon} p(\upsilon) p_k(\upsilon), \qquad (1)$$

где $p_k(v) = \sum_{\omega} p(\omega \to \upsilon) p(\omega)$ — вероятности ошибки в слове υ длины k, $p(\upsilon)$, $p(\omega)$ — вероятности встречаемости слов υ и ω.

Рассмотрим вероятности ошибок распознавания комбинированного метода $\mathfrak{R}_1$ (использован порог $p_0 = 0{,}001$), сведенные в табл. 1. Для оценки вероятностей p_{ij} для пар символ—символ было использовано тестовое множество, состоящее более чем из одного миллиона образов. Вероятности p_{ij} для перехода s_i в s_j, когда длина одного из наборов символов s_i или s_j превышает 1, оценивались экспертно.

2. Эксперименты

Оценим экспериментально вероятность словарного подтверждения несловарного распознанного слова.

Рассмотрим корпус слов русского языка, состоящий из секций с длинами, перечисленными в табл. 2.

Оценим вероятности ошибки подтверждения слова w с длиной k по формуле (1), результаты сведены в табл. 3, в которой также приводятся частоты Ne_k ошибок в словах длиной k.

На рис. 1 приведены данные табл. 3: распределения частот ошибок Ne_k и объемы секций словаря $|V_k|$ в зависимости от длины слова k. На рис. 2 изображен график вероятности p_k ошибки подтверждения в слове длины k.

Анализ ошибочных трансформаций позволяет выделить большой класс, определяемый чередованиями одной буквы в окончаниях слов. Окончания слов, которые различаются одной буквой из табл. 1, перечислены в табл. 4.

Таблица 1

Вероятности ошибок распознавания $\Re_1$

s_i	s_j	p_{ij}	s_i	s_j	p_{ij}	s_i	s_j	p_{ij}	s_i	s_j	p_{ij}	s_i	s_j	p_{ij}
д	а	0,028	й	и	0,018	ы	м	0,032	я	п	0,008	щ	ш	0,009
й	а	0,016	н	и	0,076	в	н	0,006	ч	ц	0,010	ъ	ь	0,006
л	а	0,008	п	и	0,006	и	н	0,047	м	ч	0,013	з	э	0,021
в	б	0,018	д	л	0,016	я	н	0,008	ц	ч	0,014	й	ю	0,034
ф	е	0,010	п	л	0,052	й	п	0,010	щ	ч	0,005	ыо	ью	0,02
э	з	0,014	я	л	0,008	л	п	0,011	й	ш	0,009	кж	кю	0,015

Таблица 2

Распределение длин слов русского языка

k	$\|V_k\|$	k	$\|V_k\|$	k	$\|V_k\|$	k	$\|V_k\|$	k	$\|V_k\|$
1	9	7	104 682	13	92 459	19	5789	25	159
2	86	8	144 894	14	63 911	20	3298	26	118
3	821	9	169 397	15	42 516	21	1721	27	96
4	11 345	10	174 584	16	26 783	22	963	28	79
5	31 222	11	157 312	17	16 022	23	559	29	36
6	61 858	12	126 117	18	9702	24	353	30	24

Таблица 3

Распределение вероятностей ошибки

k	$\|V_k\|$	p_k	Ne_k	k	$\|V_k\|$	p_k	Ne_k
1	9	0,0012	1	16	26 783	0,0018	1462
2	86	0,0017	7	17	16 022	0,0017	817
3	821	0,0037	158	18	9702	0,0017	496
4	11 345	0,0037	2165	19	5789	0,0016	284
5	31 222	0,0029	4248	20	3298	0,0016	157
6	61 858	0,0021	5942	21	1721	0,0016	83
7	104 682	0,0020	8676	22	963	0,0015	44
8	144 894	0,0020	11 028	23	559	0,0021	37
9	169 397	0,0020	12 393	24	353	0,0021	22
10	174 584	0,0002	12 506	25	159	0,0013	6
11	157 312	0,0002	10 590	26	118	0,0023	8
12	126 117	0,0019	7939	27	96	0,0025	7
13	92 459	0,0019	5502	28	79	0,0022	5
14	63 911	0,0018	3681	29	36	0,0019	2
15	42 516	0,0018	2395	30	24	0,0028	2

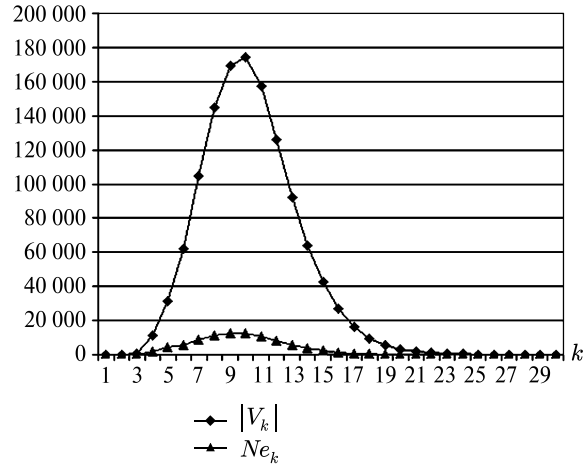

Рис. 1. Распределение частот ошибок Ne_k и объемы секций словаря $|V_k|$

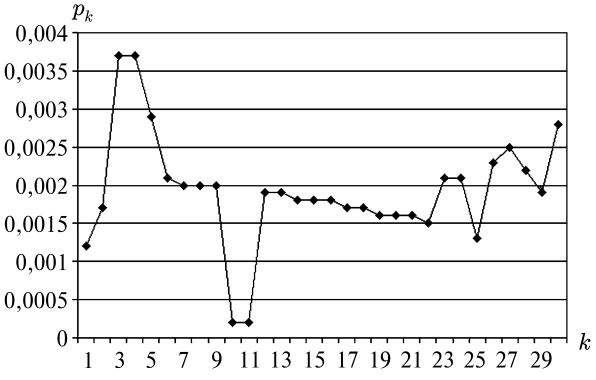

Рис. 2. Распределение вероятности p_k ошибки подтверждения в слове длины k

Таблица 4

Наиболее частые окончания, допускающие трансформации

EF_i	ET_j
ой	ою
ай	аю
ий	ии
ий	ию
уй	ую
ей	ею
яй	яю
ая	ал
яя	ял
йся	юся
нем	ием
ием	нем
вав	ван

Если при сравнении двух слов $\alpha_1\alpha_{2\ldots}\alpha_k$ и $\alpha'_1\alpha'_2\ldots\alpha'_k$ оказывается, что

$$\exists q: \alpha_{q\ldots}\alpha_k \in EF,\ \alpha'_q\ldots\alpha'_k \in ET,$$

$$\alpha_1\alpha_{2\ldots}\alpha_{q-1} = \alpha'_1\ldots\alpha'_{q-1},\ \alpha_{q\ldots}\alpha_k \neq \alpha'_q\ldots\alpha'_k,$$

то такие слова будем называть *эквивалентными*.

Будем игнорировать трансформации эквивалентных слов. Оценим вероятности ошибки подтверждения слова ω с длиной k по формуле (1) с учетом

Таблица 5

Вероятности ошибки подтверждения слова ω с длиной k

k	$\lvert V_k\rvert$	p_k	Ne_k	k	$\lvert V_k\rvert$	p_k	Ne_k
1	9	0,0012	1	16	26 783	0	1
2	86	0,0017	7	17	16 022	0	0
3	821	0,003	141	18	9702	0	0
4	11 345	0,0034	2043	19	5789	0	0
5	31 222	0,002	3361	20	3298	0	0
6	61 858	0,001	3615	21	1721	0	0
7	104 682	0,001	3933	22	963	0	0
8	144 894	0	3343	23	559	0	0
9	169 397	0	2409	24	353	0	0
10	174 584	0	1967	25	159	0	0
11	157 312	0	1203	26	118	0	0
12	126 117	0	583	27	96	0	0
13	92 459	0	250	28	79	0	0
14	63 911	0	112	29	36	0	0
15	42 516	0	25	30	24	0	0

Таблица 6

Оценки подтверждения несловарных слов

Выборка	Количество слов	Количество несловарных слов	Количество ошибок подтверждения	Вероятность ошибки подтверждения
TS_1	10 100	243	0	менее 0,01
TS_2	9054	127	1	менее 0,01
TS_3	2400	77	1	менее 0,01

отождествления эквивалентных слов, результаты сведены в табл. 5.

По сравнению с табл. 3 игнорирование трансформации эквивалентных слов приводит к повышению надежности подтверждения, например, при $k > 7$ ошибки подтверждения отсутствуют (см. рис. 3).

В целом данные табл. 3 и 5 характеризуют процедуру словарного подтверждения как способ оценки надежности распознавания при указанных требованиях к алгоритму распознавания и к распознаваемым образам слов.

В реальности распознаются произвольные слова, а не только словоформы из используемого словаря.

Оценим количество и состав несловарных слов на трех выборках документов из области книжных и математических текстов (TS_1), журнальных страниц компьютерной прессы (TS_2) и деловой переписки (TS_3). Оценки приведены в табл. 6 и свидетельствуют о том, что на указанных стендах ошибка словарного подтверждения является практически невозможным событием.

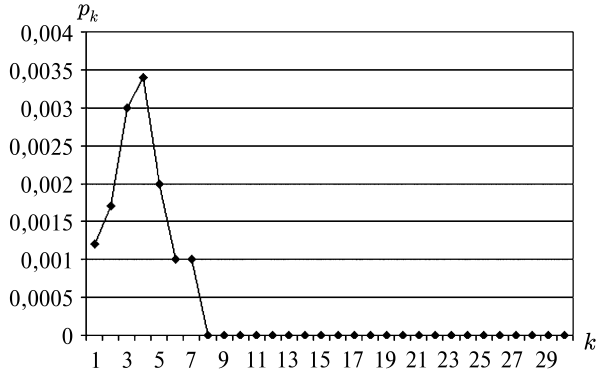

Рис. 3. Распределение вероятности p_k ошибки подтверждения в слове длины k при игнорировании окончаний

3. Практическое применение словарного подтверждения

Возможность появления ошибок при словарном подтверждении требует применения комбинирования словарных средств подтверждения надежности с другими механизмами.

Повышение надежности оценок распознавания символов в образах слов с помощью словарей, описанное в предыдущем разделе, организуется следующим образом.

Для алгоритма распознавания образов символов $\mathfrak{R}$, обладающего монотонными оценками, применяется следующий способ. Используются словари $V'_k \in V_k$, для каждой из словоформ которого $\omega = \{\omega_1, \ldots, \omega_k\} \in V'_k$ известны позиции *особых символов* $\{z_1(\omega), \ldots, z_q(\omega)\}$, $q < k$, которые могут быть распознаны ошибочно с распределением $p_{ij} = p(a_i|a_j)$, и при этом ошибочно распознанное слово подтверждается словарем.

В случае, когда распознанное слово $\omega \in V'_k$, для каждой позиции $z_i(\omega)$ определяется оценка надежности $w(z_i(\omega))$ распознавания символа, сформированная алгоритмом $\mathfrak{R}$. Для каждого из таких символов заранее должны быть рассчитаны пороги надежности $p_r(a_j)$, позволяющие принять решение о признании символа надежно распознанным на основании вероятности $p(a_j) \cdot p_r(a_j)$.

Например, для оценок надежности алгоритма $\mathfrak{R}$, превышающих 249, $p_r(a_j) < 4 \cdot 10^{-4}$ для любого из символов «абезилмнпцчшьэю». Тем самым для любого k при оценках особых символов, превышающих 249, вероятность ошибки словарного подтверждения не превышает $1{,}2 \cdot 10^{-6}$.

Другими словами, при распознавании словарных слов алгоритмом распознавания $\mathfrak{R}$ описанный способ комбинирования словарного подтверждения с монотонными оценками алгоритма делает ошибку словарного подтверждения практически невозможным событием.

Если алгоритм распознавания образов символов не обладает достаточной монотонностью оценок, также существует возможность использования словаря V, для каждого из словоформ которого известны позиции особых символов $\{z_1(\omega), \ldots, z_q(\omega)\}$. Указанные символы слова маркируются как ненадежно распознанные, либо распознаются заново алгоритмом $\mathfrak{R}$, для которого известно распределение ошибок.

Также применяется способ подтверждения надежности особых символов в слове с использованием позиционного анализа слов предложения. *Пред-*

Таблица 7

Правила для особых символов окончаний

Примеры		Различающие правила
решай, следуй, сопоставляй	решаю, следую, сопоставляю	наличие местоимений «я»/«ты» перед глаголом
акций	акции	наличие числительных, обозначающих числа больше единицы, перед существительным «акций»;
		наличие числительных «одной» или «одни» перед существительным «акции»;
		наличие глагола множественного числа после существительного «акции».
акций	акцию	наличие числительных, обозначающих числа больше единицы, перед существительным «акций»;
		наличие слов «нет», «из», «от», «без», «до» перед существительным «акций»;
		наличие числительного «одну» перед существительным «акцию».
пятилетнем	пятилетием	наличие предлогов «в», «на», «о» перед прилагательным «пятилетнем» в сочетании с последующим существительным с окончание «е»;
		наличие предлога «с», «над», «перед» перед существительным «пятилетием» в сочетании с существительным или местоимением в родительном падеже;
		наличие предлога «с» перед существительным «пятилетием» без следующего существительного;
		слово «пятилетием» может быть последним в предложении, а «пятилетнем» — нет.

Таблица 8

Ошибки подтверждения слов

Тестовая последовательность	Количество слов	Количество ошибок	Частоты ошибки подтверждения слов длиной k						
			$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$	$k>9$
TS_4	13 984	16	7	4	2	1	2	0	0
TS_5	11 036	3	1	0	1	1	0	0	0

ложением мы называем последовательности слов, удовлетворяющей следующим условиям:

- последовательность принадлежит одному фрагменту текста,
- последовательность отделена от других предложений.

К слову с особыми окончаниями прилагается набор правил, которые могут произвести выбор того или иного окончания на основе позиций слов в предложении.

В табл. 7 приведены правила для особых символов некоторых окончаний из табл. 4, часто встречаемых в тестовых последовательностях. Для других пар окончаний, таких как «ой»-«ою», «ей»-«ею», «ейся»-«еюся», «ая»-«ал», «яя»-«ял», «ав»-«ан», простых правил, основных на позиционном анализе слов в предложении, не существует. Например, приоритет одному из окончаний «ой»-«ою» может быть установлен после анализа стиля текстового фрагмента, а одного предложения для такого анализа недостаточно.

В реальности кроме рассмотренных в описанной выше модели трансформаций слов возможны другие случаи ошибок при словарном подтверждении, например, при распознавании образа, не являющегося образом слова, или распознавании образа слова, напечатанного на языке, отличном от алфавита распознавания.

Рассмотрим несколько наборов слов TS_4, TS_5, подтвержденных словарным механизмом в программе распознавания текстовых документов OCR Cuneiform [1], и оценим количество ошибок подтверждения. Результаты сведены в табл. 8.

Использование оценок распознавания алгоритма $\mathcal{A}$ с порогом, равным 249, и правил позиционного анализа обеспечивает отсутствие ошибок при подтверждении слов.

Выводы

Предложенный способ словарного подтверждения распознанных символов обеспечивает с избытком потребности адаптивного распознавания в

классах документов с хорошим и средним качеством печати. Описанный способ внедрен в программу распознавания печатных документов OCR Cuneiform [3, 1].

Литература

1. OCR Cuneiform, [Электронный ресурс], http://cognitiveforms.ru/products/cuneiform/
2. *Арлазаров В. Л., Логинов А. С., Славин О. А.* Характеристики программ оптического распознавания текста // Программирование. 2002. № 3. С. 45–63.
3. *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.
4. *Кочин Д. Ю., Хлебутин П. С.* Разработка многомодульных программных комплексов // Сб. трудов ИСА РАН «Развитие безбумажной технологии в организационных системах». М.: URSS, 1999. С. 110–126.

Славин Олег Анатольевич. Заведующий лабораторией ИСА РАН, д. т. н. Окончил Московский институт радиотехники, электроники и автоматики в 1988 г. Количество печатных работ: 68, в том числе 1 монография. Область научных интересов: распознавание образов, искусственный интеллект, моделирование электромагнитных процессов. E-mail: oslavin@cs.isa.ru

Янишевский Игорь Михайлович. Старший научный сотрудник ИСА РАН, к. ф.-м. н., окончил МГУ им. М. В. Ломоносова (ф-т ВМиК) в 1988 г. Количество печатных работ: 16. Область научных интересов: теория случайных процессов, распознавание образов, теория оптимального управления. E-mail: igor_y@cs.isa.ru