V Всероссийский симпозиум
"Инфраструктура научных информационных ресурсов и систем"

Построение системы массового распознавания архивных документов с автоматической корректировкой результатов

СПб ГУП «Санкт-Петербургский информационно-аналитический центр»
Смирнов Сергей Владимирович

Санкт-Петербург,
2015 г.

АКТУАЛЬНОСТЬ

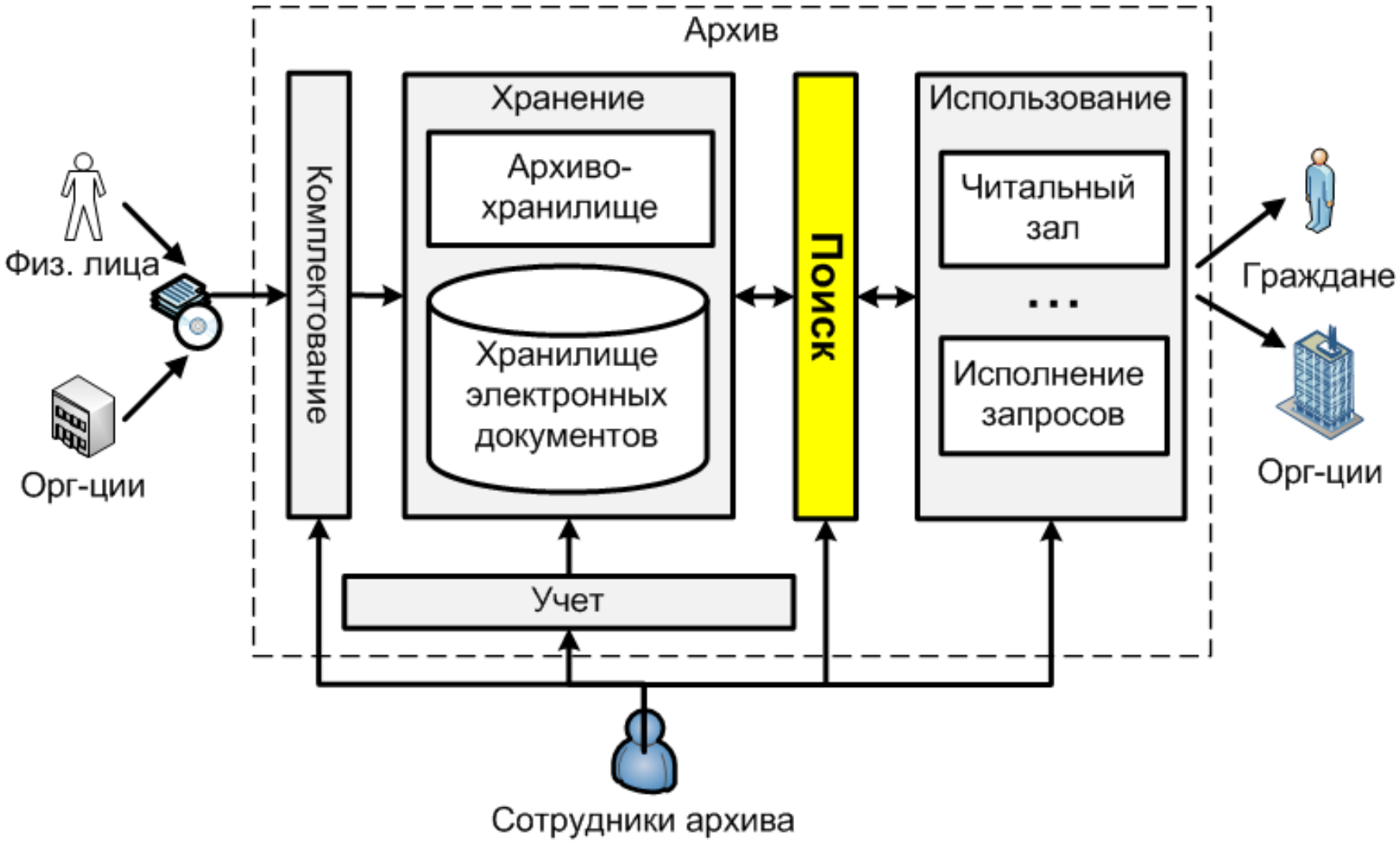
«Информационное общество РФ (2011-2020 годы)»:

- оцифровка объектов культурного наследия, включая архивные фонды;
- развитие средств обработки и предоставления удаленного доступа к цифровому контенту.

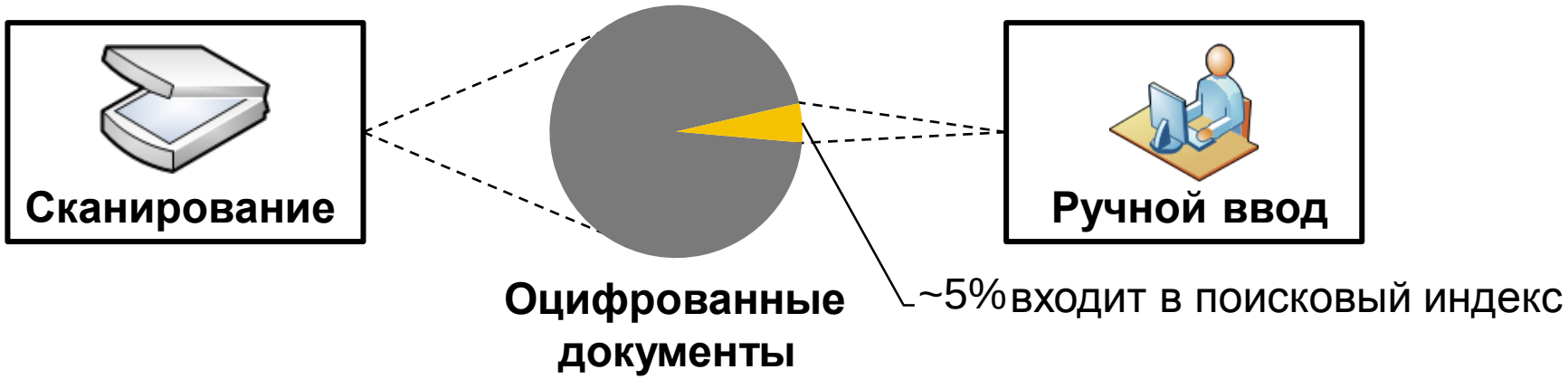
Концепция развития архивного дела в РФ до 2020 года

- оцифровка основных информационно-поисковых средств государственных и муниципальных архивов;
- представление онлайн доступа к ним и к тематическим базам данных.

ПРЕДМЕТНАЯ ОБЛАСТЬ



ПРОБЛЕМАТИКА

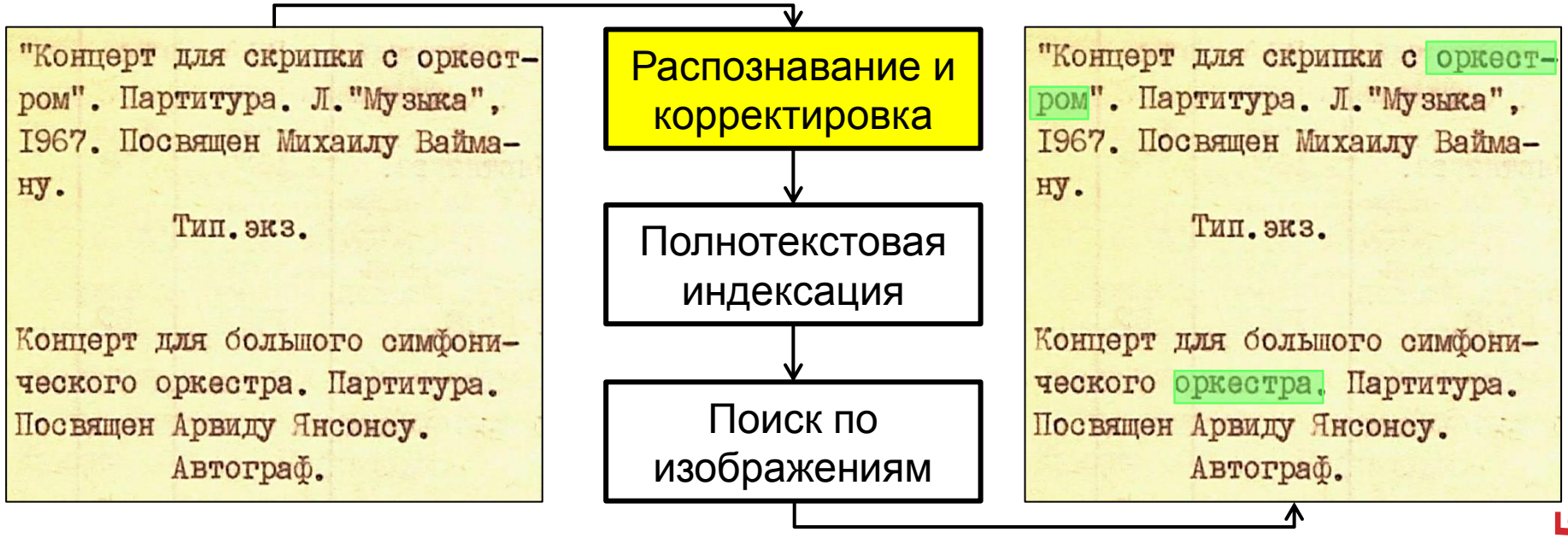


- **Низкая эффективность ручного ввода поисковых реквизитов**
  - Высокая трудоемкость и стоимость ручного труда
  - Низкая скорость пополнения поисковой базы
- $V_{\text{сканирование}} \gg V_{\text{ручной ввод}}$
- **Низкий процент поискового покрытия базы электронных документов**

ЦЕЛЬ РАБОТЫ

Разработка технологии и системы распознавания архивных документов с автоматическим обнаружением и корректировкой допущенных ошибок.

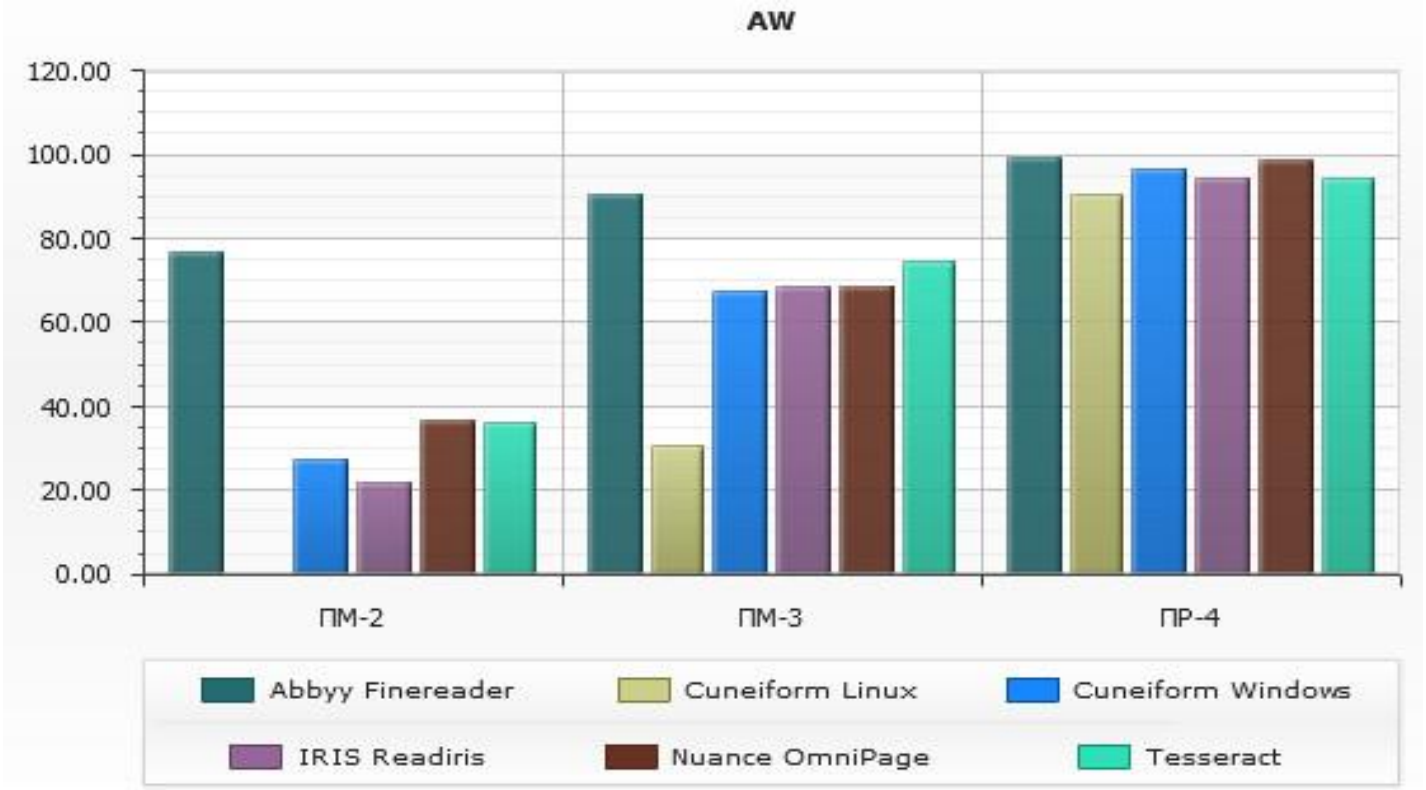
- Обработка документов разнообразных тематических областей
- Обработка узкоспециализированной терминологии
- Обеспечение автоматической работы
- Подготовка результатов для системы поиска по изображениям



СРАВНИТЕЛЬНЫЙ АНАЛИЗ OCR СИСТЕМ

Точность распознавания на уровне слов

Набор изображений	Max	Min
печатная машинка, среднее качество ПМ-2	76,80%	**22,01%**
печатная машинка, высокое качество ПМ-3	90,55%	30,28%
принтер, очень высокое качество ПР-4	**99,25%**	90,20%



АНАЛИЗ СУЩЕСТВУЮЩИХ ПОДХОДОВ К КОРРЕКТИРОВКЕ

Из существующих методов были использованы:

- ❑ Алгоритм нахождения минимального расстояния между словами (расстояние Левенштейна*)
- ❑ Алгоритм поиска схожих слов методом анаграмм**

Особенности:

- ✓ Не требуют предварительного обучения
- ✓ Могут применяться для обработки текстов на любых языках

***** Левенштейн В. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академий Наук СССР. 1965. Т. 163. № 4. С. 845-848

** Reynaert M. Non-interactive OCR Post-correction for Giga-Scale Digitization Projects // Computational Linguistics and Intelligent Text Processing. 2008. pp. 617–630

МЕТОД АНАГРАММ

Ключевое слово:

$w = \text{слого}$

Словарь:

$\text{W} = \{\ldots, \text{слово}, \ldots\}$

Ключевой алфавит

$$\begin{Bmatrix} \text{с}, \\ \ldots, \\ \text{г}, \\ \ldots, \\ \text{го} \end{Bmatrix}$$

Ключевой хэш-алфавит

$$\begin{Bmatrix} hash(\text{с}), \\ \ldots, \\ hash(\text{г}), \\ \ldots, \\ hash(\text{го}), \\ 0 \end{Bmatrix}$$

Поисковый алфавит

$$\begin{Bmatrix} ' ', \\ \ldots, \\ \text{в}, \\ \ldots, \\ \text{ая} \end{Bmatrix}$$

Поисковый хэш-алфавит

$$\begin{Bmatrix} hash(' '), \\ \ldots, \\ hash(\text{в}), \\ \ldots, \\ hash(\text{ая}), \\ 0 \end{Bmatrix}$$

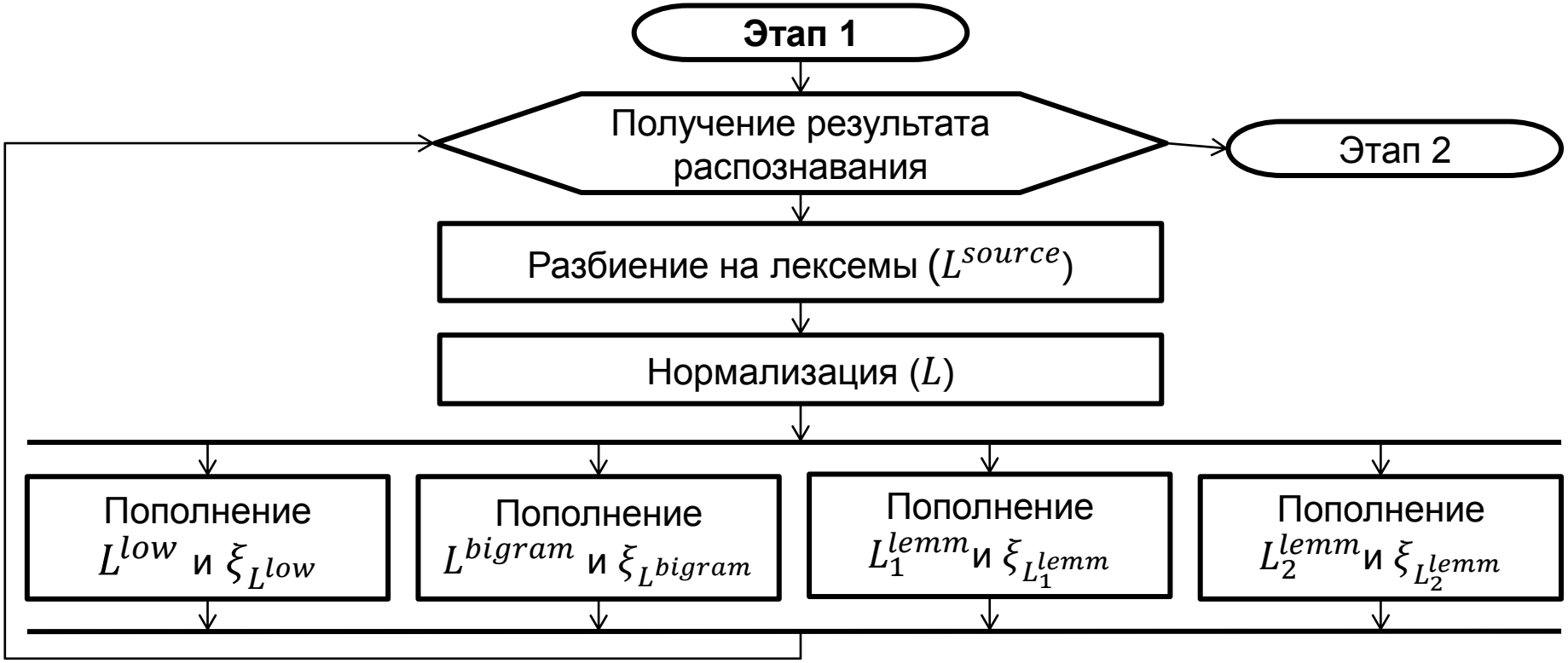
$$hash(\text{слого}) - hash(\text{г}) + hash(\text{в}) = hash(\text{слово}) = hash(\text{волос})$$

$$hash(w) = \sum_{i=1}^{|w|} int(c_i)^n = \sum_{i=1}^{|w|} hash(c_i)$$

$w = \{c_1 \ldots c_{|w|}\}$, $int(c)$ — значение символа c в кодовой таблице UTF-8.

МЕТОД АВТОМАТИЧЕСКОЙ КОРРЕКТИРОВКИ РЕЗУЛЬТАТОВ РАСПОЗНАВАНИЯ НА ОСНОВЕ РЕЙТИНГО-РАНГОВОЙ МОДЕЛИ ТЕКСТА

1. ПОДГОТОВКА СТРУКТУР ДАННЫХ



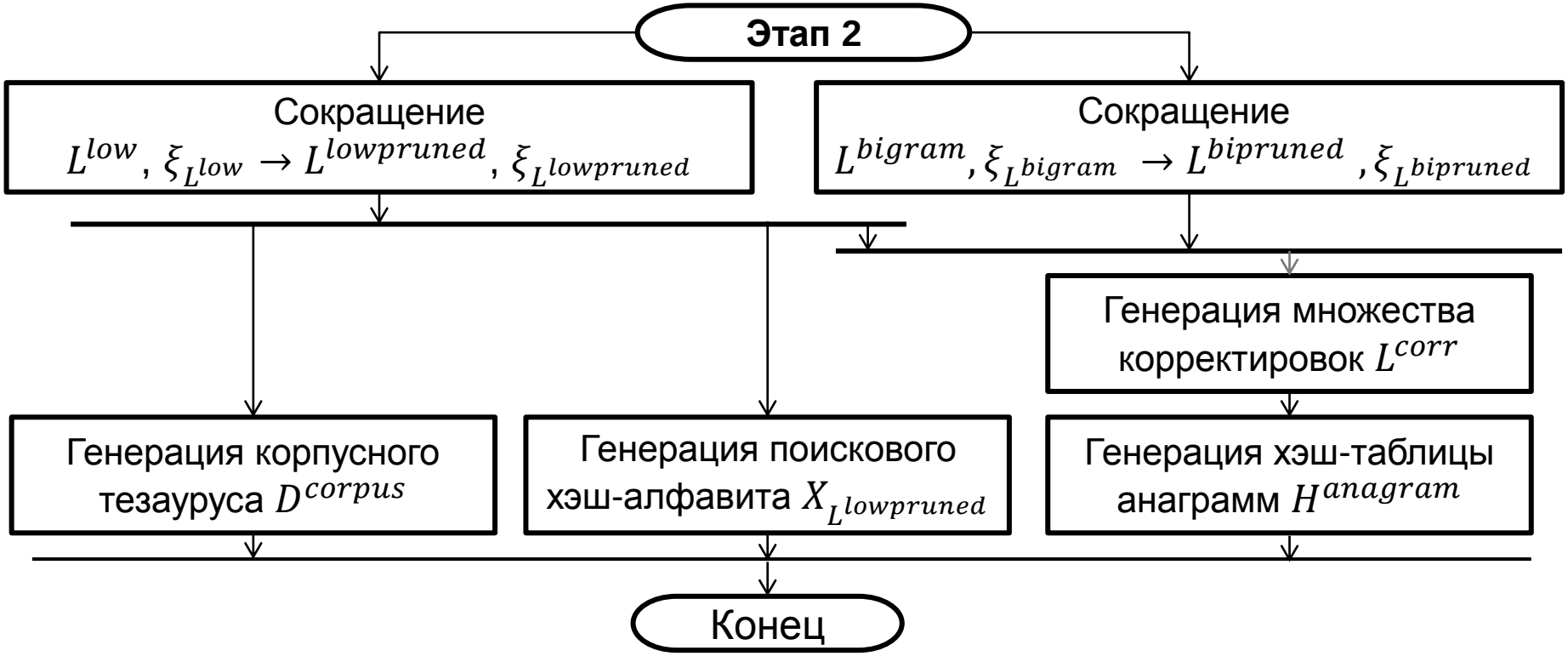
$$L^{low} = \{lower(s) |\ s \in L\},\ \ \xi_{L^{low}} = \{< s, fr > | s \in L^{low}\},\ \ fr \geq 1$$

$$L^{bigram} = \{(lower(s_1), lower(s_2)) | s_1, s_2 \in L, seq(s_1, s_2) \vee seq(s_2, s_1) = 1\}$$

$$L_1^{lemm} = \{morph(s) | s \in L, s \notin D_{stop}\}, L_2^{lemm} = \{(b_1, b_2) | b_1, b_2 \in L_1^{lemm}, seq(b_1, b_2) = 1\}$$

$lower(s)$ - перевод строки в нижний регистр, $seq(a_1, \dots, a_z)$ - последовательность элементов от a_1 до a_z, $morph(s)$ - нормальные формы слова s.

1. ПОДГОТОВКА СТРУКТУР ДАННЫХ



$$L^{corr} = L^{lowpruned} \cup \{concat(s_1, " ", s_2) | (s_1, s_2) \in L^{bipruned}\}$$

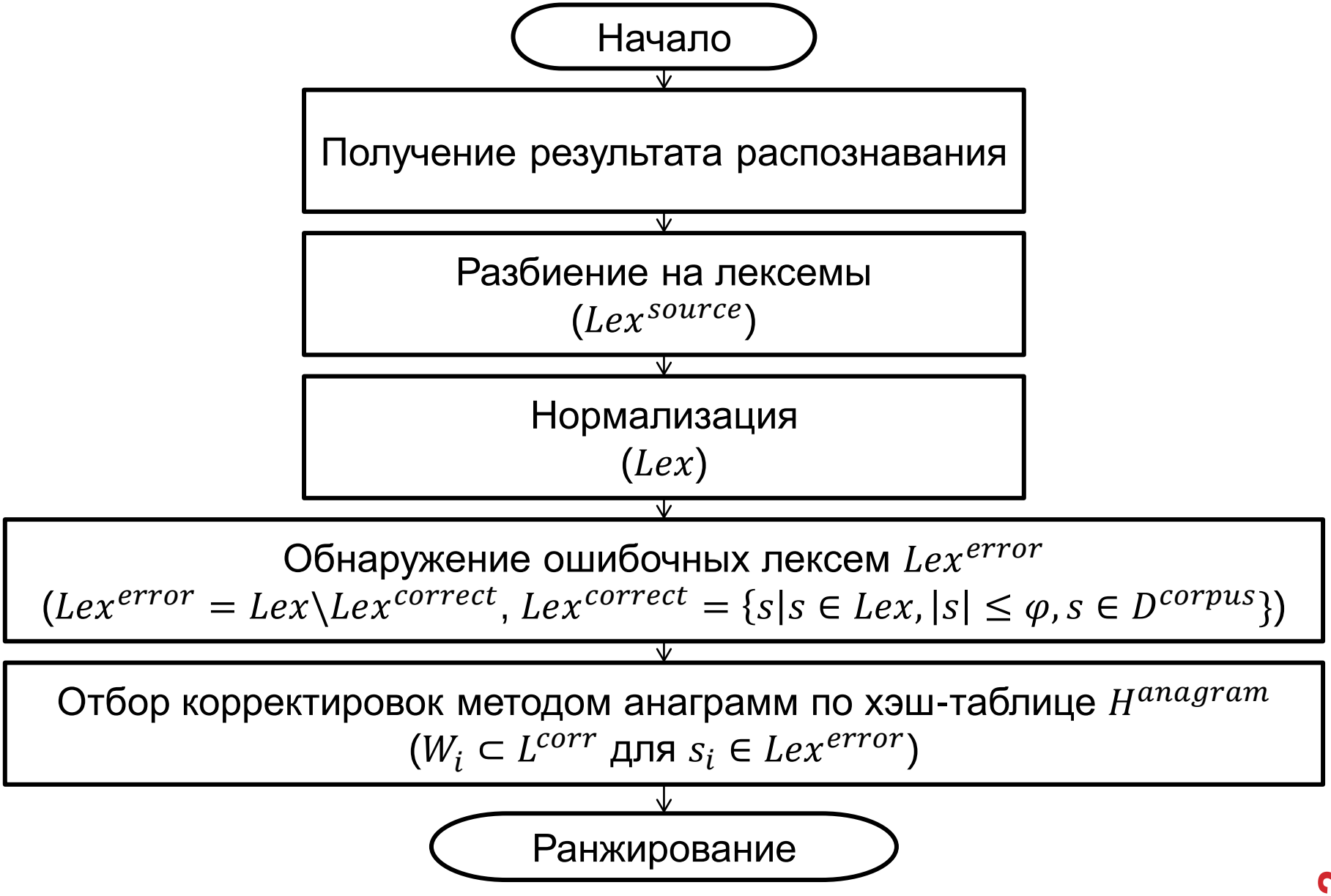
$$X_{L^{lowpruned}} = \{hash(ngrams(\{w\})) \mid w \in L^{lowpruned}\} \cap ngrams(\{а..я\} \cup \{-, " "\})$$

$$H^{anagram} = \{< hash(s), < s, \xi_{L^{corr}}(s) > > | s \in L^{corr}\}$$

$$D^{corpus} = L^{lowpruned} \cap (D^{gen} \cup D^{spec}),$$

где D^{gen} — словарь общих слов русского языка, D^{spec} — тематические тезаурусы

2. ГЕНЕРАЦИЯ КОРРЕКТИРОВОК



3. РАНЖИРОВАНИЕ КОРРЕКТИРОВОК

$$W \xrightarrow{1} \vec{W} \xrightarrow{2} \widehat{W}$$

Шаг 1. Инвариантная оценка корректировки

$$score(s,w) = \ln(\xi_{L^{corr}}(w)) \times \left(|w| - LD(s,w)\right) \times r(w) \times d_{factor},$$

$$d_{factor} = \begin{cases} 3, & \text{если } w \in D^{corpus} \\ 1, & \text{если } w \notin D^{corpus} \end{cases}$$

где $\xi_{L^{corr}}(w)$ — частота повторения слова w во всем корпусе слов L^{corr}, $|w|$ — длина слова w, $r(w)$— количество повторений корректировки w в ходе отбора методом анаграмм.

Шаг 2. Вычисление финального ранга

$$Rank(s,w) = \frac{score(s,w)}{\sum_{j=1}^{|\vec{W}|} score(s,w_j)} \times P(w),$$

$$P(w_i^k) = \frac{\sum_{j=1}^{|\vec{W}_{i-1}|} \xi_{L_2^{lemm}}(morph(w_{i-1}^j), morph(w_i^k))}{\sum_{j=1}^{|\vec{W}_{i-1}|} \xi_{L_1^{lemm}}(morph(w_{i-1}^j))}$$

4. ФОРМИРОВАНИЕ РЕЗУЛЬТАТА

Результат распознавания:

$$RES = \left\{ < s, w^{best}, W^{alternate} > | \; s \in Lex \right\}$$

$w^{best} \in \widehat{W}$ - наилучшая корректировка,

$W^{alternate} = \widehat{W} \backslash \{w^{best}\}$ - множество дополнительных корректировок.

Правила выбора наилучшей корректировки:

1) $w^{best} = \underset{w \in (\widehat{W} \cap D^{abbr})}{Arg\max} Rank(s, w)$ *- аббревиатуры*

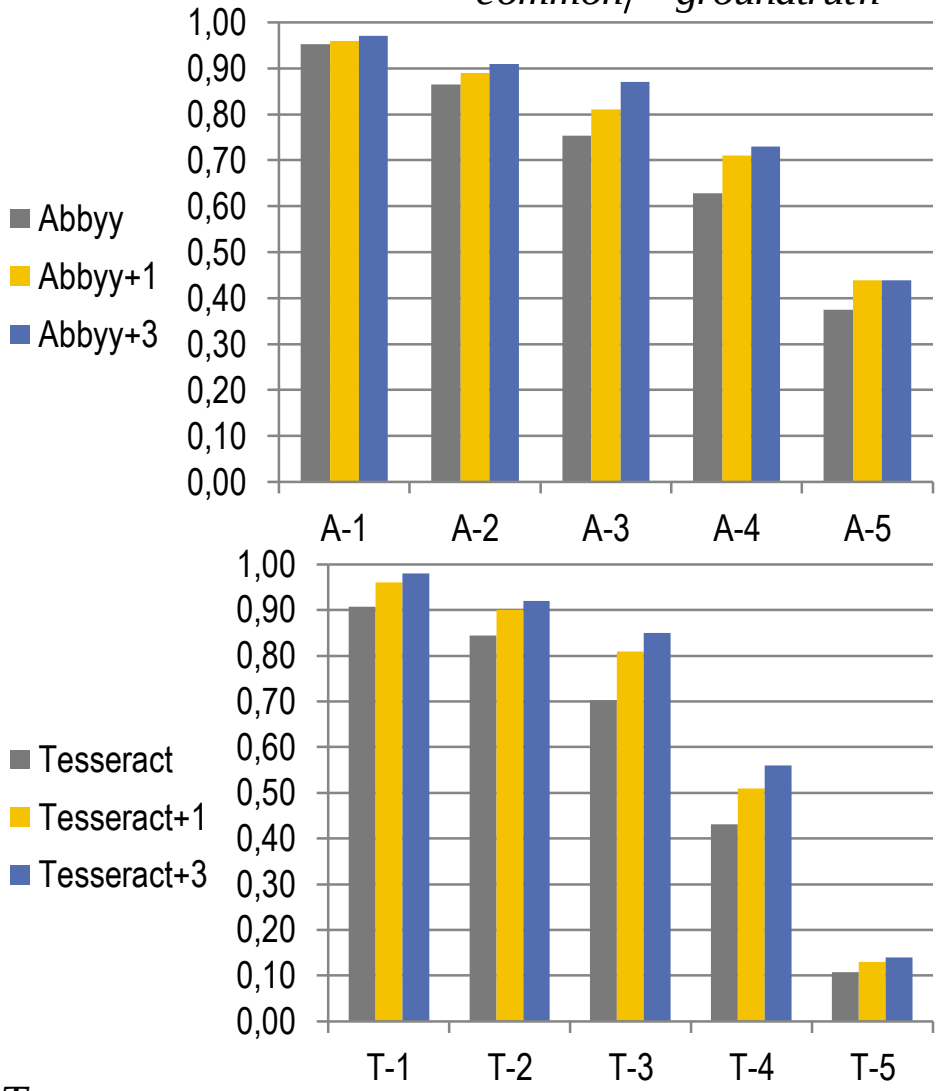
2) $w^{best} = \underset{w \in (\widehat{W} \cap (D^{surname} \cup D^{name}))}{Arg\max} Rank(s, w)$ *- имена собственные*

3) $w^{best} = \widehat{w}_1, \widehat{W} = \{\widehat{w}_1 \ldots \widehat{w}_{|\widehat{W}|}\}$ *- с наивысшим рангом*

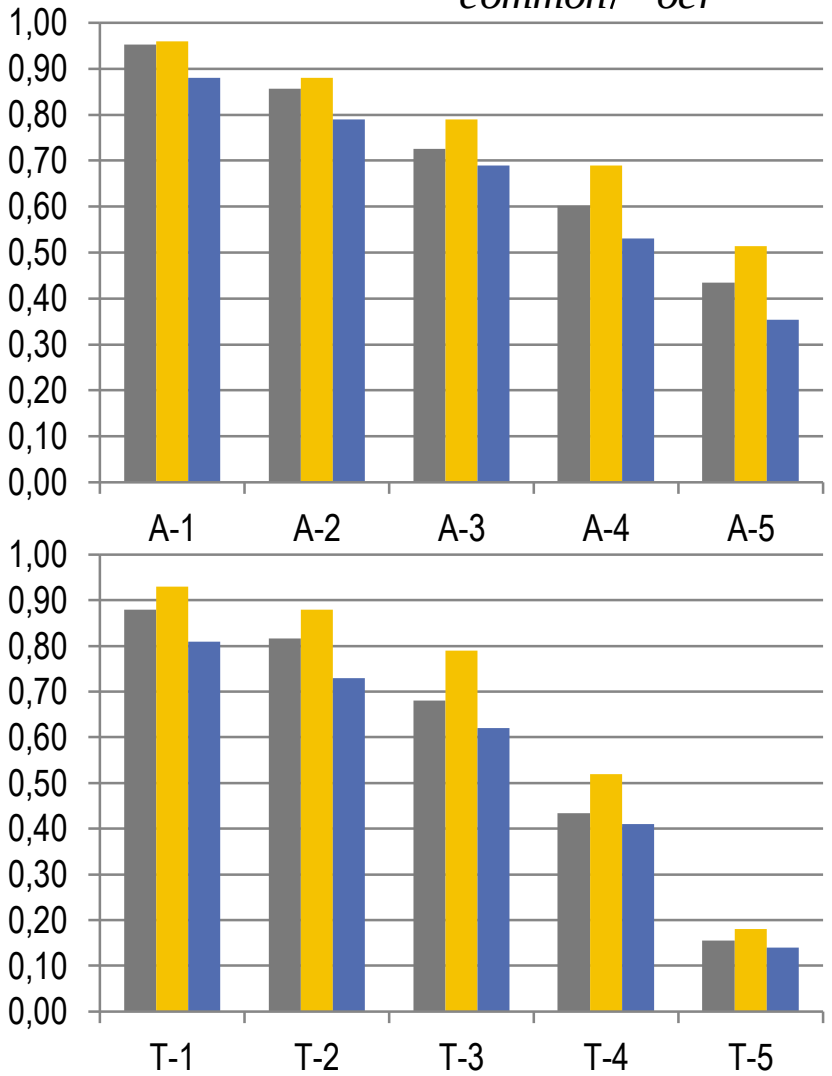
ОЦЕНКА МЕТОДА

приращение до +15%

$Recall = T_{common}/T_{groundtruth}$



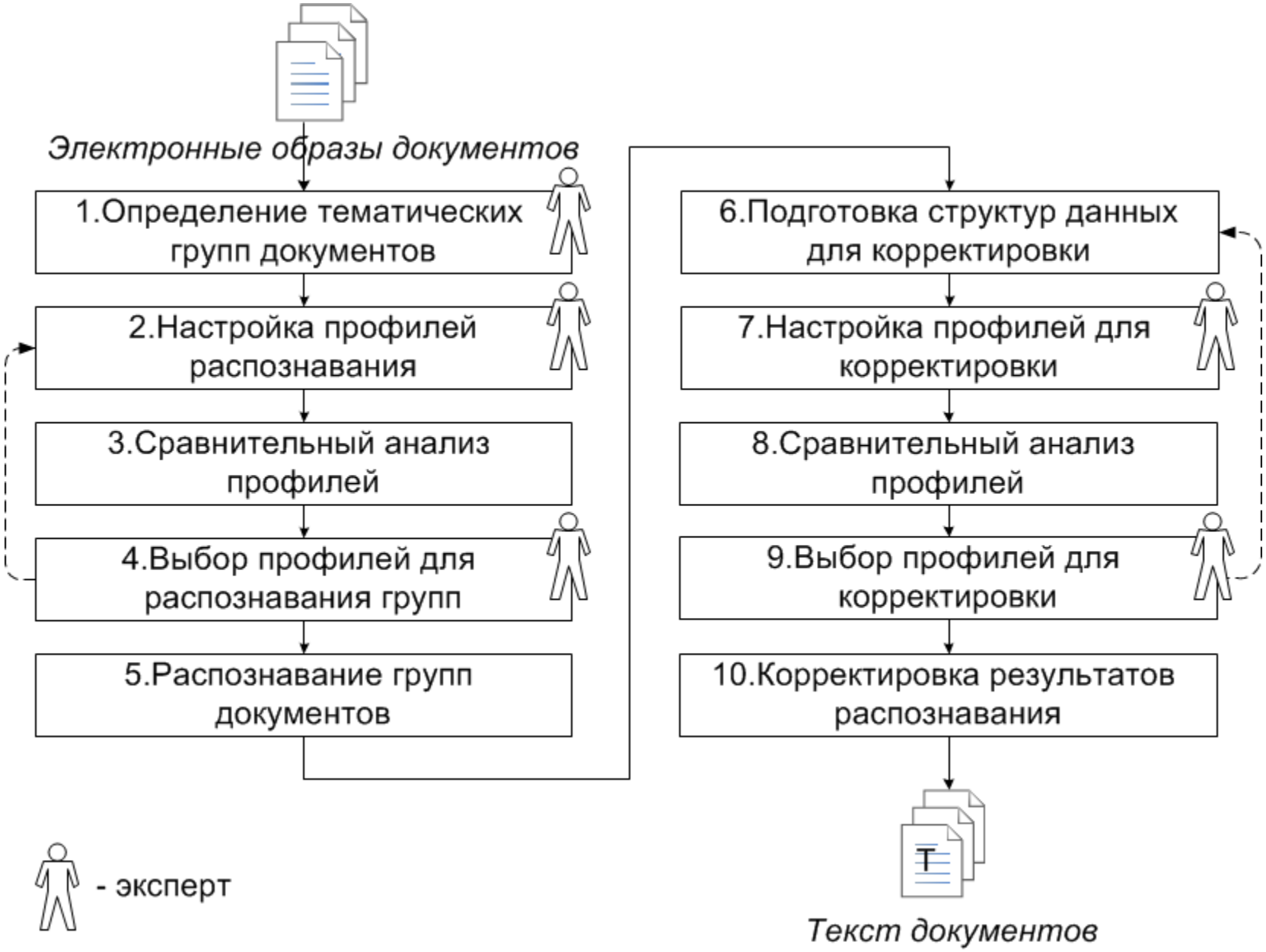
$Precision = T_{common}/T_{ocr}$



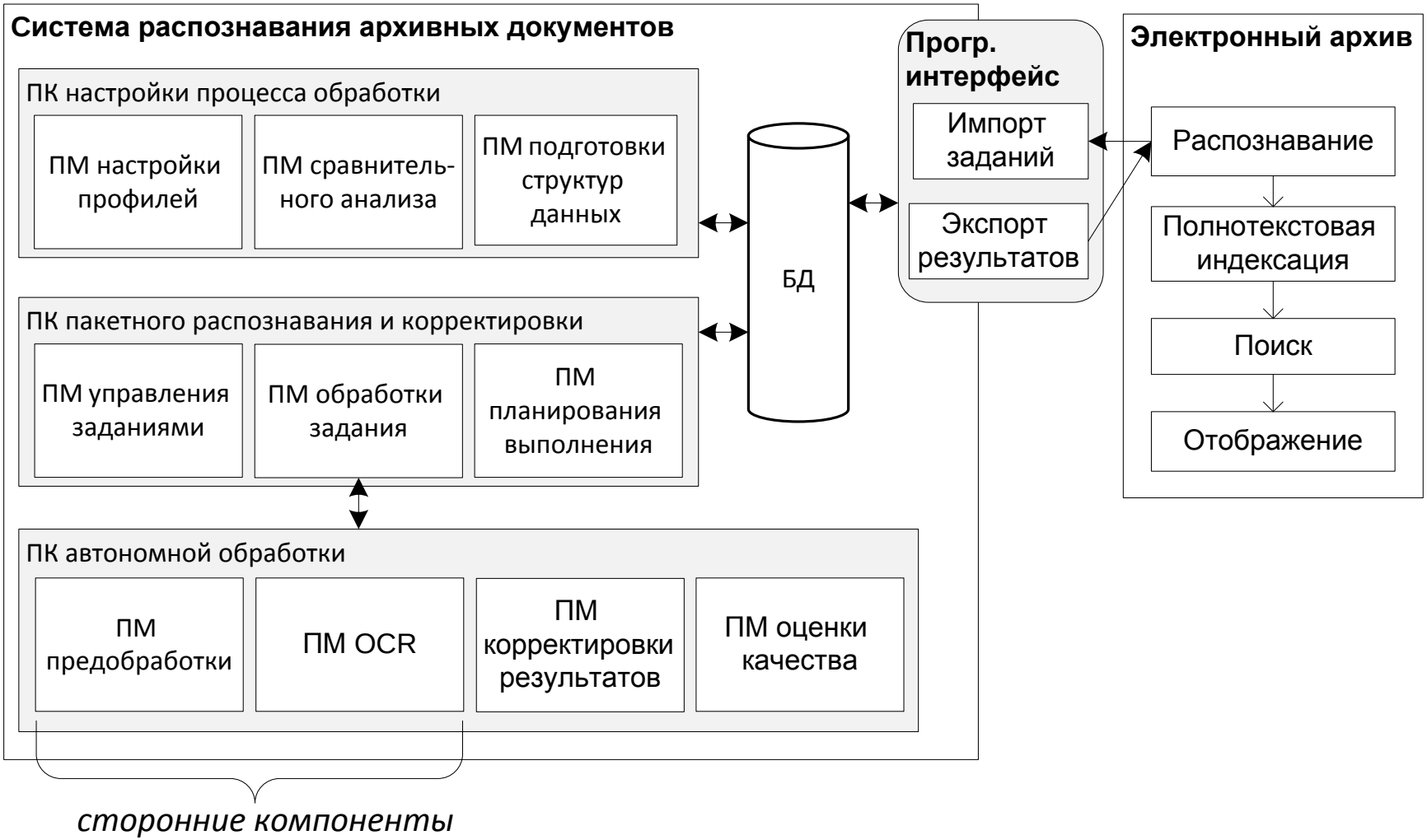
T_{common} — количество поисковых токенов эталона, содержащихся в распознанном тексте;

$T_{groundtruth}$ — количество токенов в эталоне; T_{ocr} — количество токенов в результате распознавания.

ТЕХНОЛОГИЯ ОБРАБОТКИ ДОКУМЕНТОВ

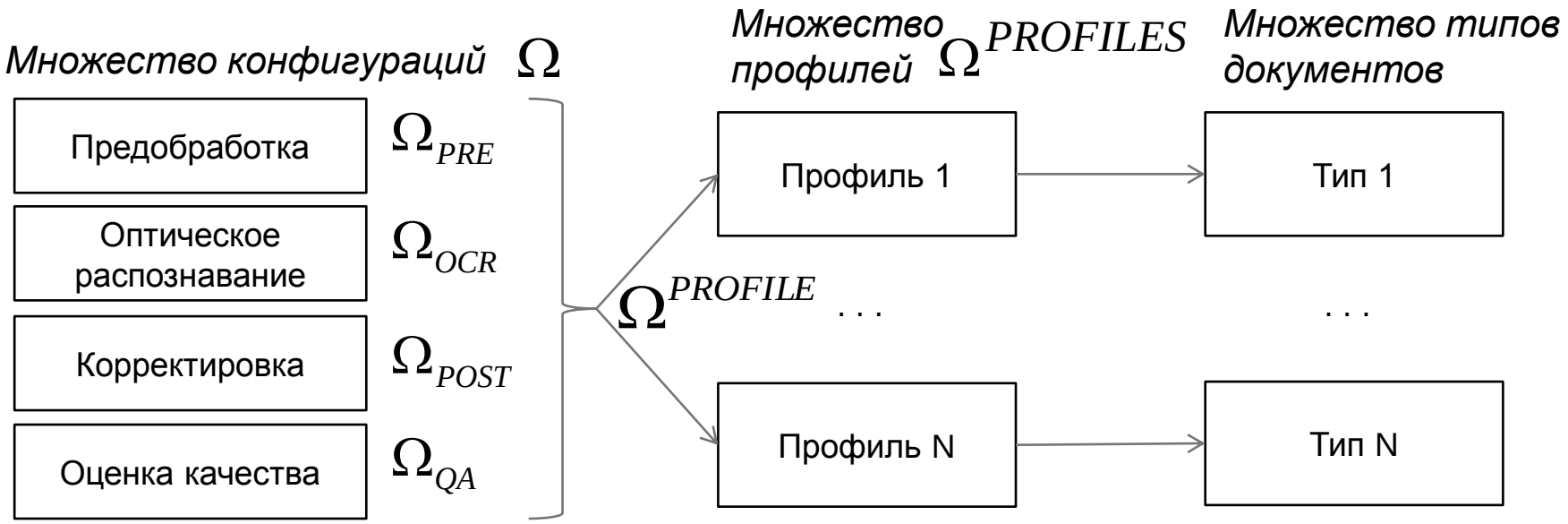


КОМПОНЕНТНАЯ МОДЕЛЬ СИСТЕМЫ



ПК – программный комплекс, ПМ – программный модуль, БД – база данных

ИНСТРУМЕНТАРИЙ КОНФИГУРИРОВАНИЯ

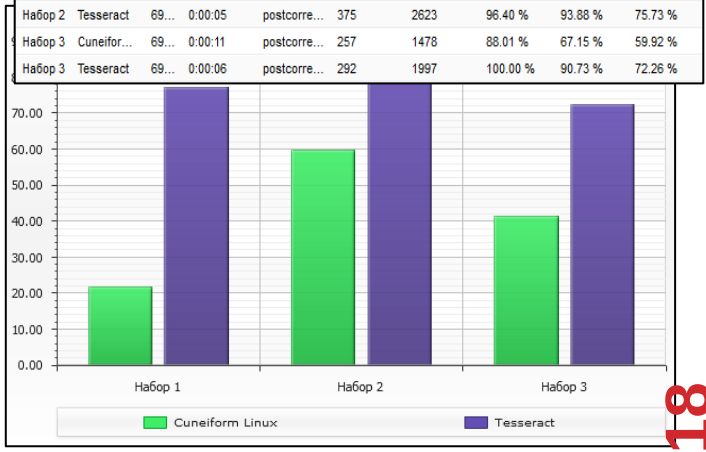


Критерии оценки качества распознавания

T_C	кол-во символов в эталоне	A_{W_b}	точность в словах без учета порядка
r_{T_C}	коэффициент распознанных символов	P	точность наличия поисковых токенов
T_W	кол-во слов в эталоне	R	полнота наличия поисковых токенов
r_{T_W}	коэффициент распознанных слов	F	F-мера наличия поисковых токенов
A_D	словарная точность	P_{alt}	точность с корректировками
A_C	точность в символах	R_{alt}	полнота с корректировками
A_W	точность в словах	F_{alt}	F-мера с корректировками

Сравнительный анализ профилей

Набор...	Профайл	За...	tOCR	stage	Tw	Tc	rTw	rTc	AD
Набор 1	Cuneifor...	69...	0:00:06	postcorre...	185	954	69.29 %	52.88 %	64.32 %
Набор 1	Tesseract	69...	0:00:04	postcorre...	226	1630	84.64 %	90.35 %	80.97 %
Набор 2	Cuneifor...	69...	0:00:08	postcorre...	358	2208	92.03 %	79.03 %	67.32 %
Набор 2	Tesseract	69...	0:00:05	postcorre...	375	2623	96.40 %	93.88 %	75.73 %
Набор 3	Cuneifor...	69...	0:00:11	postcorre...	257	1478	88.01 %	67.15 %	59.92 %
Набор 3	Tesseract	69...	0:00:06	postcorre...	292	1997	100.00 %	90.73 %	72.26 %



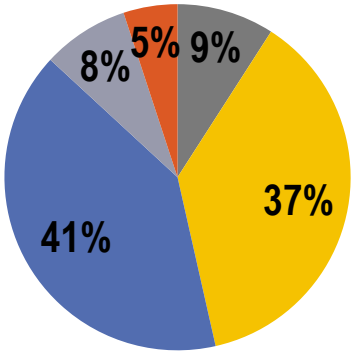
РЕЗУЛЬТАТЫ ОПЫТНОЙ ЭКСПЛУАТАЦИИ

Результаты распознавания и корректировки всего корпуса документов

Документы		Кол-во образов	Распознано лексем	Кол-во ошибок	Исправлено ошибок	Словарная точность		
архив	тип					до	после	Δ+
ЦГА	описи	342 319	46 267 201	17 706 507	36%	62%	75%	13%
	указатели	8 251	1 289 212	625 751	24%	51%	63%	12%
ЦГАИПД	указатели	207 254	22 094 096	10 647 194	52%	52%	77%	25%
ЦГАЛИ	описи	28 251	3 200 646	1 091 795	33%	66%	77%	11%
ЦГАЛС	описи	59 064	9 353 017	4 529 334	**80%**	52%	90%	**38%**
ЦГАНТД	описи	63 524	5 837 216	1 854 795	34%	68%	79%	11%
Итого:		**708 663**	**88 041 388**	**36 455 376**	**46%**	**59%**	**77%**	**18%**

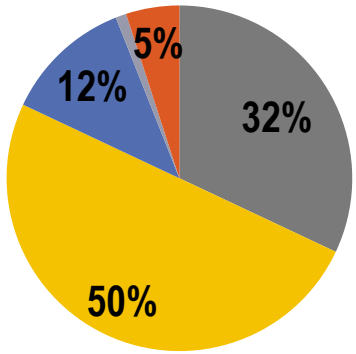
Распределение количества образов по диапазонам словарной точности

до корректировки



Диапазоны словарной точности A_D :

- 80%-100%
- 60%-80%
- 40%-60%
- 20%-40%
- 0%-20%



после корректировки

Выводы

- Распознано **32 608** документов (**708 663** изображения), получено **88 млн** лексем (слов).
- Количество ошибочных слов (не входящих в словарь) сократилось на **46%**, исправлено **16 497 948** ошибочных слов, значение словарной точности увеличилось на **18%**.
- Экономия **~50** человеко-лет ручного труда по вводу данных.

РЕЗУЛЬТАТЫ РАБОТЫ

1. Разработан метод автоматической корректировки ошибок распознавания архивных документов на основе рейтинго-ранговой модели текста, производящий поиск корректировок по тезаурусам, предварительно извлеченным из результатов распознавания и текстов одной тематической области.
2. Разработаны правила ранжирования и выбора наилучших корректировок, основанные на вычислении инвариантной оценки соответствия и вероятности нахождения финального слова n-граммы по известным предыдущим словам.
3. Разработан инструментарий, позволяющий эксперту производить настройку системы для обработки архивных документов различных тематических областей путем выбора оптимального набора параметров на основе сравнительного анализа качества распознавания тестовых изображений.
4. Разработаны технология и система распознавания архивных документов и автоматической корректировки результатов, успешно интегрированные с системой электронного архива и позволяющие производить массовую параллельную обработку документов в пакетном режиме.

СПАСИБО ЗА ВНИМАНИЕ!