# Японский: буквенные n-граммы для распознавания

Куликов Алексей Владимирович

Московский физико-технический институт (государственный университет)
Факультет инновация и высоких технологий
Кафедра компьютерной лингвистики

Научный руководитель — А.И. Андрианов

Москва, 2017

# Содержание

# Постановка задачи

**Цель работы** – сравнить эффективность различных символьных $n$-граммных моделей в задаче исправления ошибок OCR в японском языке.
Из цели работы вытекают следующие **задачи:**

- Рассмотреть существующие подходы к $n$-граммному моделированию японского языка;
- Реализовать некоторые символьные $n$-граммные модели;
- Развернуть систему для тестирования и сравнения моделей.

# Формальная постановка задачи

Формализуем задачу:

- *Алфавит* $\Sigma = \{a, b, c, ..\}$.
- *Текст* $Text \in \Sigma^+$ делится на конечное число *предложений* $S = \{S_1, S_2, S_3, ...\} : Text = S_1 S_2 S_3 ....$
- Для каждого предложения $S$ существует $k$ вариантов: $S_{set} = \{S_{etalon}, S_{test_1}, ..., S_{test_{k-1}}\}$.
- *Оценивающий алгоритм (estimator)* $\Theta : S \to \mathbb{R}^+$.
- Среди $k$ вариантов предложения выбирается наилучший: $S_{best} = \underset{S_{set}}{\operatorname{argmax}} \Theta(S)$.
- *Качество алгоритма* $Q(\Theta) = \dfrac{\#\{\textit{угаданных предложений}\}}{\#\{\textit{всего предложений}\}}$.

Необходимо сравнить различные алгоритмы по качеству.

- $N$-граммные с фиксированным $n$, $n \in \{1, 2, 3\}$
- Backoff-модель, $n_{max} \in \{3, 5, 7\}$

$$P_n(w_i|w_{i-n+1}, ..., w_{i-1}) = \begin{cases} C(w_i|w_{i-n+1}, ..., w_{i-1}) & \text{if } C(w_i|w_{i-n+1}, ..., w_{i-1}) > k \\ P_{n-1}(w_i|w_{i-n+2}, ..., w_{i-1}) & \text{otherwise} \end{cases}$$

- Модель Катца (Katz), $n_{max} \in \{3, 5, 7\}$

$$P_n\left(w_i|w_{i-n+1}...w_{i-1}\right) = \begin{cases} d_{w_{i-n+1}...w_i} \dfrac{C(w_{i-n+1}...w_i)}{C(w_{i-n+1}...w_{i-1})} & \text{if } C(w_{i-n+1}...w_i) > k \\ \alpha_{w_{i-n+1}...w_{i-1}} P_{n-1}\left(w_i|w_{i-n+2}...w_{i-1}\right) & \text{otherwise} \end{cases}$$

# Корпус

HTML-страницы с различных японских сайтов, 163244 штуки.
Статистики по корпусу после его обработки:

| Параметр | Значение |
|---|---|
| Размер (kB) | 1 721 504 |
| Символов | 640 604 961 |
| среди них уникальных | 6 861 |
| Предложений | 79 497 345 |

Таблица 1: Характеристики корпуса

# Количество униграмм

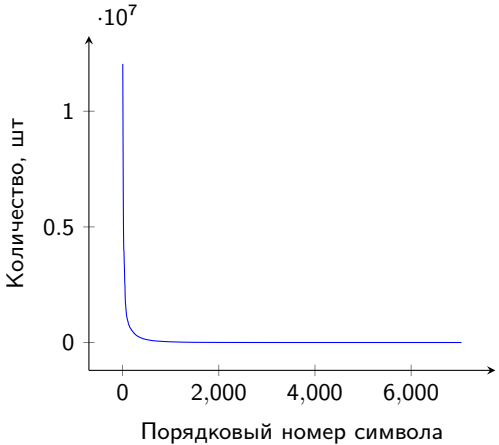



Рис. 1: Распределение частот униграмм

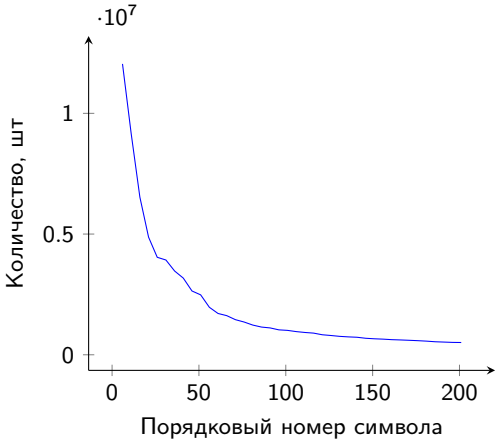
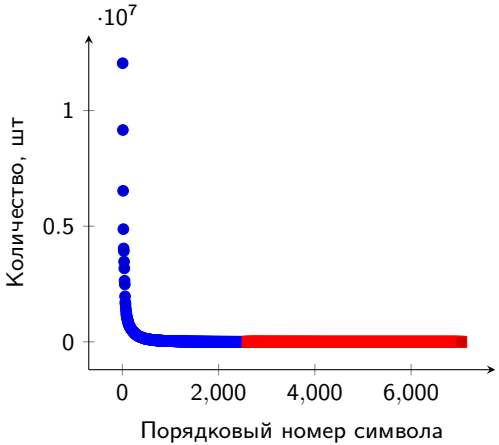



Рис. 2: Частоты униграмм – голова распределения

# Buckets

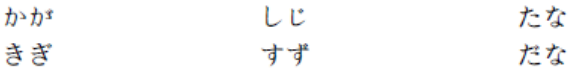
*Корзина (бакет, bucket)* – множество символов, которые считаются статистически малозначимыми и заменяются на U+FFFD (Unicode Replacement Character).
Выбран бакет с $|\Sigma_{B^*}| = 4800$.



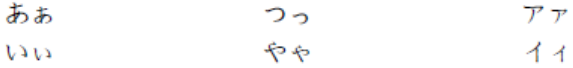
# Шум как эмуляция ошибок OCR

Необходимо эмулировать ошибки OCR.

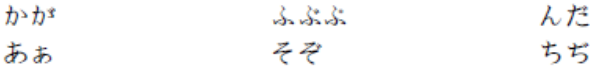
- **KaGa**




- **BigSmall**




- **Mix**



# Пример шума

| Шум | Текст |
|---|---|
| Эталон | キャンペーンは終了致しました。 |
| KaGa | ギャンペーン**ば**終了致しました。 |
| BigSmall | キ**ヤ**ンペーンは終了致しました。 |
| Mix | **ギ**ャンペーンは終了致しました。 |

# Распределение шумов по униграммам

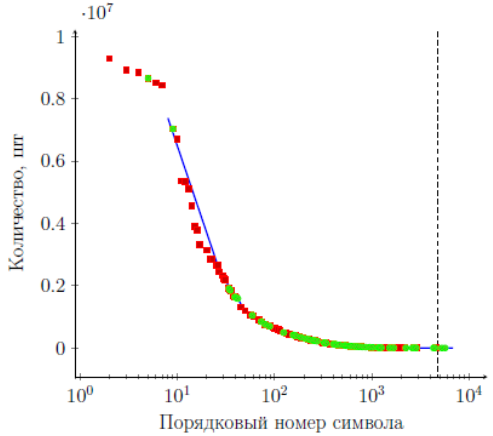



Рис. 13: Распределение частот шума **Mix**

Униграммная модель.
Результаты показаны для основного бакета, $|\Sigma_{B^*}| = 4800$.

| Шум | Оценка модели |
|---|---|
| **KaGa** | 0.75 |
| **BigSmall** | 0.88 |
| **Mix** | 0.76 |

Таблица 2: Baseline

# Результаты

| $M \backslash N$ | **KaGa** | **BigSmall** | **Mix** |
|---|---|---|---|
| Ngram(1) (Baseline) | 0.75 | 0.88 | 0.76 |
| Backoff(5) | 0.89 | 0.93 | 0.90 |
| Katz(5) | 0.961 | 0.965 | 0.962 |

Таблица 3: Результаты эксперимента: accuracy

| $C \backslash N$ | KaGa | BigSmall | Mix |
|---|---|---|---|
| 0.9 | 0.86 | 0.94 | 0.86 |
| 0.95 | 0.979 | 0.988 | 0.981 |
| 0.97 | 0.966 | 0.986 | 0.967 |
| 0.99 | 0.94 | 0.98 | 0.95 |

Таблица 4: *

Accuracy

| $C \backslash N$ | KaGa | BigSmall | Mix |
|---|---|---|---|
| 0.9 | 0.53 | 0.43 | 0.52 |
| 0.95 | 0.61 | 0.69 | 0.62 |
| 0.97 | 0.77 | 0.85 | 0.77 |
| 0.99 | 0.91 | 0.94 | 0.91 |

Таблица 5: *

Доля классифицированных

Таблица 6: Результаты эксперимента с *Confidence*

Выбрана модель Katz($n = 5, C = 0.97$).

# Статистика по *n*-граммам

| Symbol | TP | FN | FP | TN | Precision | Recall | F1 | Accuracy |
|---|---|---|---|---|---|---|---|---|
| な | 20554 | 185 | 802 | 127893 | 0.962446151 | 0.991079608 | 0.976553035 | 0.993395077 |
| だ | 789 | 25 | 50 | 9853 | 0.940405244 | 0.969287469 | 0.954627949 | 0.993001773 |
| ビ | 1300 | 80 | 10 | 306 | 0.992366412 | 0.942028986 | 0.966542751 | 0.946933962 |
| ピ | 403 | 6 | 67 | 712 | 0.857446809 | 0.985330073 | 0.916951081 | 0.938552189 |
| ミ | 481 | 12 | 19 | 1229 | 0.962 | 0.975659229 | 0.96878147 | 0.982194141 |
| マ | 1229 | 19 | 12 | 481 | 0.990330379 | 0.984775641 | 0.987545199 | 0.982194141 |
| び | 43633 | 9 | 59 | 117 | 0.998649638 | 0.999793777 | 0.99922138 | 0.998448126 |
| ひ | 229 | 63 | 5 | 21435 | 0.978632479 | 0.784246575 | 0.870722433 | 0.996870974 |
| の | 497614 | 9348 | 10 | 6194 | 0.999979905 | 0.981560748 | 0.99068472 | 0.981764185 |
| め | 6194 | 10 | 9348 | 497614 | 0.398533007 | 0.998388137 | 0.569667985 | 0.981764185 |
| さ | 10916 | 24 | 0 | 228 | 1 | 0.997806216 | 0.998901903 | 0.997851003 |
| ぎ | 228 | 0 | 24 | 10916 | 0.904761905 | 1 | 0.95 | 0.997851003 |
| で | 42499 | 944 | 4 | 6954 | 0.999905889 | 0.978270377 | 0.988969818 | 0.981190849 |
| て | 6954 | 4 | 944 | 42499 | 0.88047607 | 0.999425122 | 0.936187399 | 0.981190849 |
| に | 249152 | 1310 | 94 | 6978 | 0.999622863 | 0.994769666 | 0.997190359 | 0.994548293 |
| ぬ | 52 | 0 | 541 | 123404 | 0.087689713 | 1 | 0.16124031 | 0.995636991 |
| ス | 15085 | 220 | 11 | 206 | 0.99927133 | 0.985625613 | 0.992401566 | 0.985117897 |
| ヌ | 170 | 0 | 99 | 3765 | 0.63197026 | 1 | 0.774487472 | 0.975458602 |
| シ | 6379 | 152 | 122 | 10915 | 0.981233656 | 0.976726382 | 0.978974831 | 0.984403461 |

| Модель | Место на диске | RAM |
|---|---|---|
| 1-gram | 100 KB | $\approx$ 1 MB |
| 3-gram trie | 178 MB | $\approx$ 7 MB |
| 5-gram trie | 1.5 GB | $\approx$ 40 GB |
| 7-gram trie | 3.7 GB | $\approx$ 110 GB |

Таблица 7: Затраты ресурсов, память, $|\Sigma| = 4800$

| Модель | 10000 примеров | 1000000 примеров |
|---|---|---|
| 1-gram | $<$ 1 мин | $\approx$ 20 мин |
| 3-gram trie | $\approx$ 1 мин | $\approx$ 1.5 ч |
| 5-gram trie | $\approx$ 2 мин | $\approx$ 3 ч |

Таблица 8: Затраты ресурсов, время, $|\Sigma| = 4800$

# Сравнение с другими результатами

Работа Nagata, использующая backoff *n*-граммы для исправления ошибок OCR.
Также использует размеченный корпус со словным делением.

| Модель | Accuracy |
|---|---|
| Katz | 0.96–0.986 |
| Nagata | 0.97–0.98 |

Таблица 9: Сравнение с результатами Nagata

Нехватка информации в *n*-граммной модели.
Существуют примеры, которые модель незаслуженно трактует как неверные.

| Тип | Предложение | Оценка |
|---|---|---|
| Эталон | ご要望**な**わせて、勉強会を開催。 | 56 |
| Шум | ご要望**に**わせて、勉強会を開催。 | 63 |

Таблица 12: Пример ошибочной классификации

# Использованное ПО

- Python 3.5.2 (pickle, dot, nltk, unicodedammit)
- Bash
- Graphviz

Спасибо