

# Японский: буквенные n-граммы для распознавания

## Контроль НИР

Куликов А.В., гр. 397  
*Руководитель: Андрианов А.И.*

ABBYY-MIPT

Москва, 2017

- Japanese kanji/kana OCR.
- Существуют путающиеся символы, например:

お и ん

- *Цель работы:* построить и сравнить различные эвристики для исправления ошибок OCR, используя буквенную n-граммную модель японского языка.

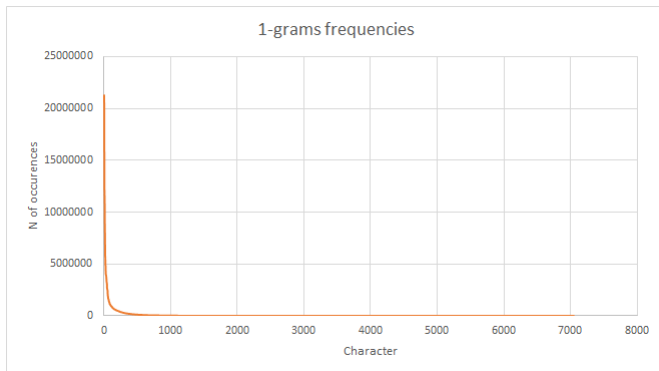
- 1 Обработка корпуса, buckets
- 2 Сбор статистики по n-граммам
- 3 Шум как имитация ошибок OCR
- 4 Модели оценки зашумлённого текста
- 5 Результаты работы моделей
- 6 Эксперимент

# Обработка корпуса

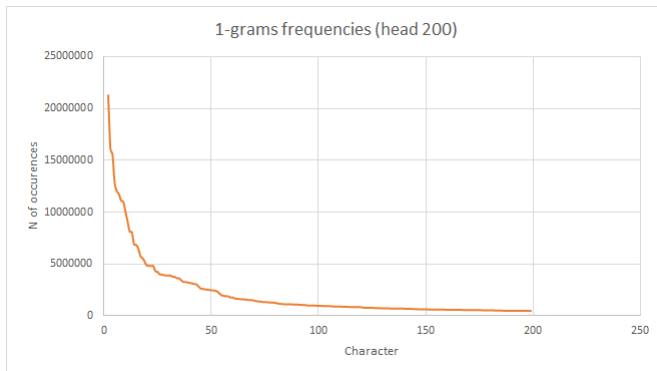
Частоты символов в корпусе (всего 5.6E+8).

Размер алфавита  $|\Sigma| = 7047$ .

Это 33% от Unicode CJK диапазона.

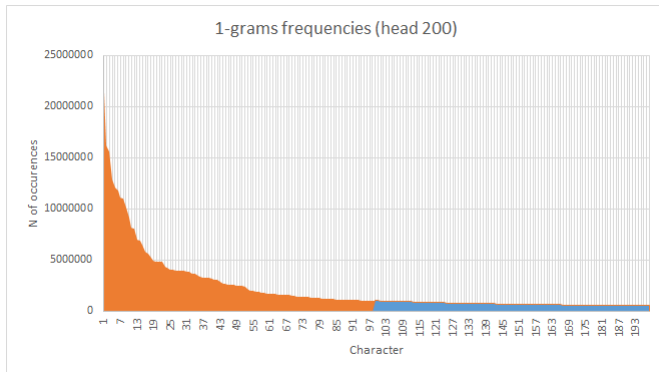


Голова этого графика.



Отбросим хвост с частотами  $\leq 1000$   
( $< 0,12\%$  корпуса,  $|\Sigma^*| = 2600$ ).

Заменяем эти символы на  $\diamond$ .



Границу можно изменять.

# Сбор статистики по n-граммам

- Храним статистику в боре (trie).
- Библиотека `pygtrie` даёт удобный функционал.
- $n_{max} = 7$ .
- Размер бора 5GB.
- Учёт не встреченных символов:  
Good-Turing,  $\diamond$ .

n	Различных n-грамм
1	4430
2	482607
3	3436987
4	10025503
5	19051342
6	28679559
7	37723274



# Шум как имитация ошибок OCR

- Ошибки OCR имитируются искусственным случайным зашумлением текста.
- Рандом зафиксирован.
- Новые списки путающихся символов:
  - Ka-ga (диакритика) (KaGa);

か<sub>и</sub> が

- Halfwidth-fullwidth forms (HFW);

サ<sub>и</sub> サ

- Микс (Mix).

- 2-gram, 3-gram;
- Katz backoff (KBO)  $n = 7$ ;

$$P_{b0}(w_i | w_{i-n+1} \dots w_{i-1}) = \begin{cases} d_{w_{i-n+1} \dots w_i} \frac{C(w_{i-n+1} \dots w_i)}{C(w_{i-n+1} \dots w_{i-1})} & \text{if } C(w_{i-n+1} \dots w_i) > k \\ \alpha_{w_{i-n+1} \dots w_{i-1}} P_{b0}(w_i | w_{i-n+2} \dots w_{i-1}) & \text{otherwise} \end{cases}$$

- Stupid backoff (SBO)  $n = 7$ .

$$d_{w_{i-n+1} \dots w_i} = \text{const}$$

$$\alpha_{w_{i-n+1} \dots w_{i-1}} = \text{const}$$

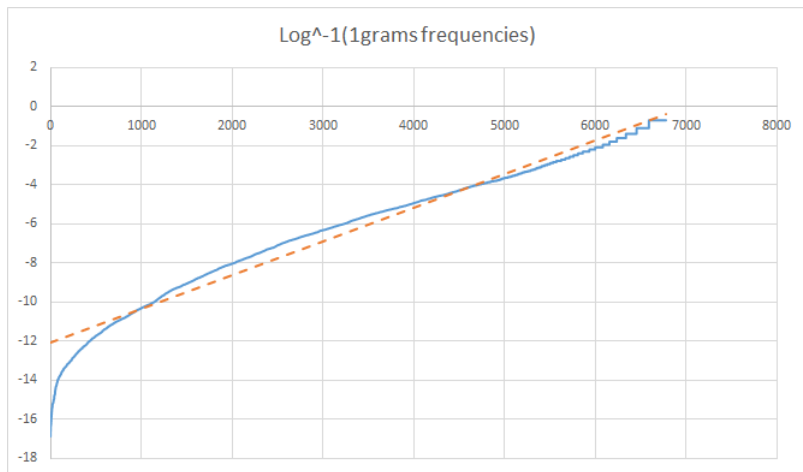
# Результаты работы моделей

Число – процент угаданных предложений в тексте.

Модель/Шум	KaGa	HFW	Mix
3-gram	82.3%	85.6%	83.7%
Stupid BO	87.8%	90.2%	89.2%
Katz BO	93.1%	94.9%	93.9%

В BO моделях есть возможность дальнейшей доводки параметров.

# Эксперимент с Zipf



- Как размер корпуса влияет на хвост распределения?
- Как мы можем имитировать большой корпус?
- Попробуем стохастически его уменьшать и посмотреть на графики.

- Тоньше настроить модель Katz BackOff.
- Попробовать имитировать большой корпус.
- Лучше визуализировать результаты.
- Писать текст.

- Foundations of Statistical Natural Language Processing /  
C. D. Manning, H. Schutze.
- Efficient In-memory Data Structures for N-Grams Indexing /  
D. Robenek, J. Platos, V. Snasel

# Спасибо