Японский: буквенные n-граммы для распознавания

Контроль НИР

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

ABBYY-MIPT

Москва, 2017

Задача

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

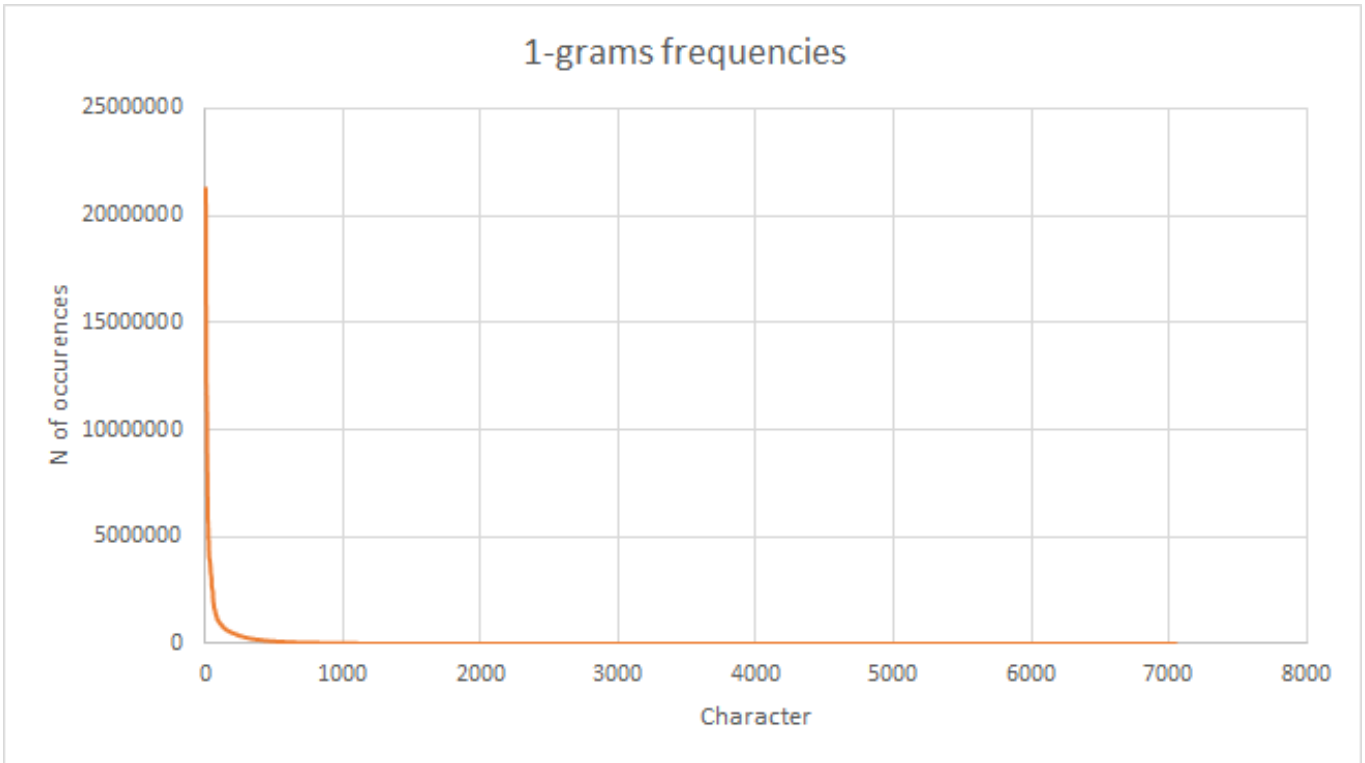
- *Цель работы:* построить и сравнить различные эвристики для исправления ошибок OCR, используя буквенную n-граммную модель японского языка.

Overview

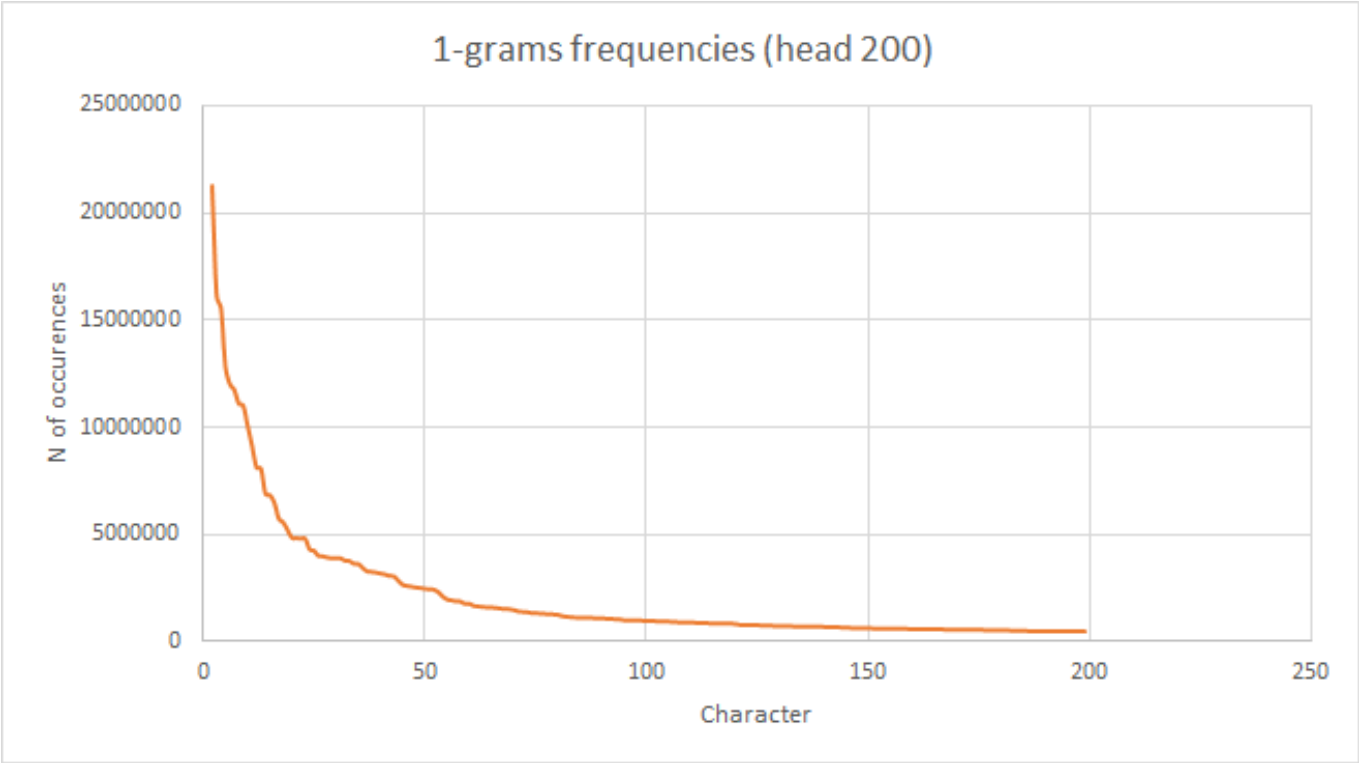
1. Обработка корпуса, buckets
2. Сбор статистики по n-граммам
3. Шум как имитация ошибок OCR
4. Модели оценки зашумлённого текста
5. Результаты работы моделей
6. Эксперимент

Обработка корпуса

Частоты символов в корпусе (всего 5.6E+8).
Размер алфавита $|\Sigma| = 7047$.
Это 33% от Unicode CJK диапазона.

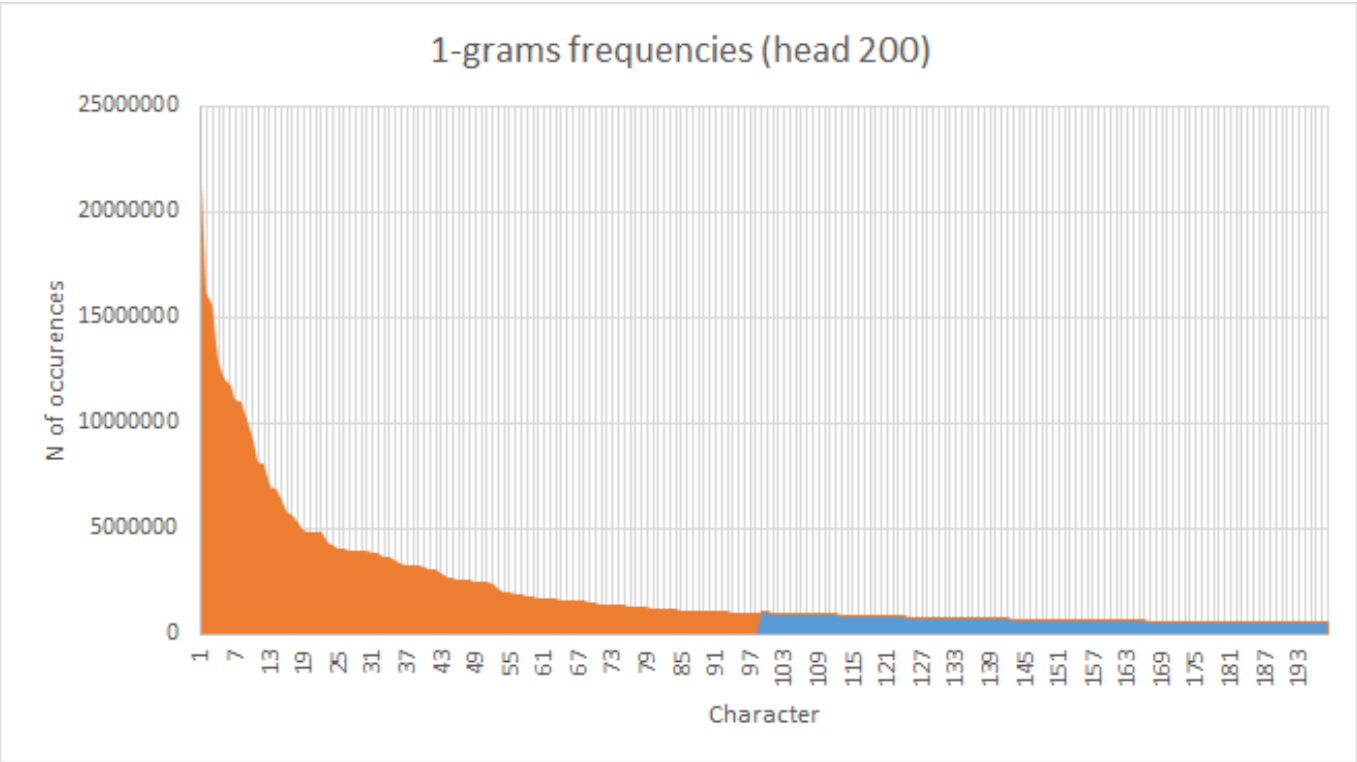


Голова этого графика.



Buckets

Отбросим хвост с частотами ≤ 1000 ($< 0,12\%$ корпуса, $|\Sigma^*| = 2600$).
Заменим эти символы на $\blacksquare$.



Границу можно изменять.

Сбор статистики по n-граммам

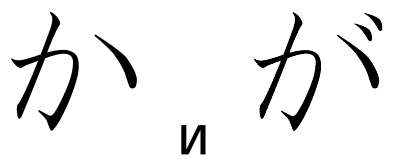
- Храним статистику в боре (trie).
- Библиотека pygtrie даёт удобный функционал.
- $n_{max} = 7$.
- Размер бора 5GB.
- Учёт невстреченных символов:
  Good-Turing, .

Trie stats

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);



  - Halfwidth-fullwidth forms (HFW);

  - Микс (Mix).

Модели оценки

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

$$P_{b0}\left(w_i \middle| w_{i-n+1} \dots w_{i-1}\right) =$$

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

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

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

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

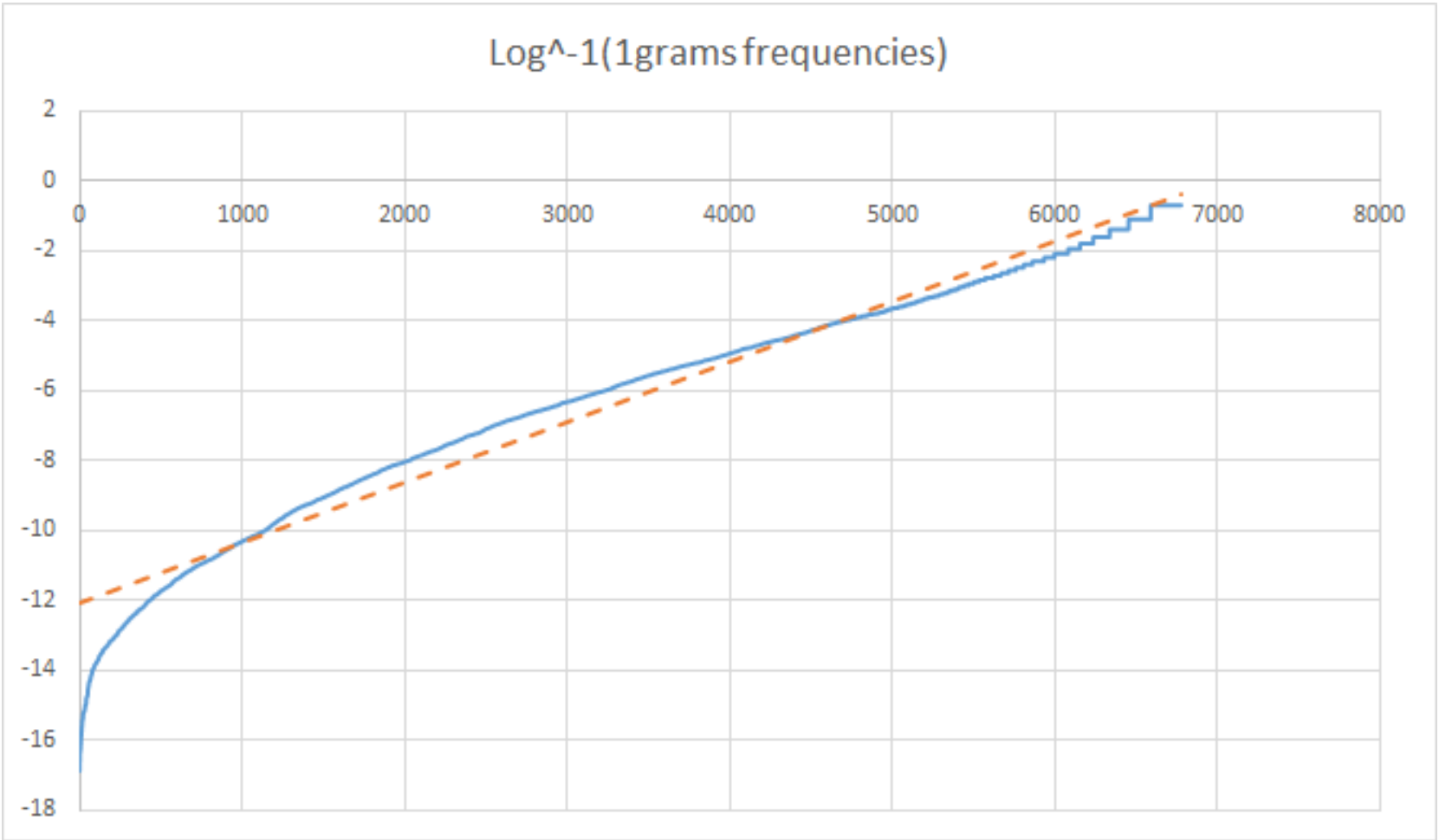
Результаты работы моделей

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



Эксперимент с Zipf

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

Планы на будущее

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

Список литературы

- 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