Японский: буквенные n-граммы для распознавания

Контроль НИР

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

ABBYY-MIPT

Москва, 2016

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.

Текущие результаты

- Был выбран, получен и адаптирован корпус;
- Были получены статистики по n-граммам;
- Был разработан настраиваемый алгоритм для зашумления корпуса;
- Был определён Baseline;
- Были получены первые результаты.

Корпус

- Корпус html-страниц с различных сайтов, доступный компании ABBYY;
- Был приведён к plain-utf-8 представлению;
- Итоговый размер $\approx$ 1.5 GB;
- Был разделён на 3 неравных подкорпуса (debug/train/test).

Корпус – исходные кодировки

Документов	Кодировка
68789	utf-8
46870	iso-8859-2
42015	shift_jis
2562	euc-jp
1575	cp932
544	ascii
436	windows-1253
256	iso-8859-7
$\approx 5\%$	ещё 6 штук

Всего около 160k документов.

Корпус – статистики по n-граммам

- Получены по train-подкорпусу ($\approx 400MB$):

n	bins	outcomes	size ($\approx$)	time ($\approx$)
1-gram	5188	2283229	100 KB	2 mins
2-gram	402035	5426594	6.8 MB	50 mins
3-gram	2455307	10407170	48.7 MB	2 hrs

- Сериализованы в pickle-dump nltk.FreqDist (весьма эффективно по памяти).

Пошумим

- Список частых OCR-ошибок:

```
[う]
5=410
ろ=115
勺=175

[え]
之=2069
九=2008
大=3138
无=1688
```

- Различные стратегии замены (абсолютное/относительное значение, какой символ подставлять и т.д.).

Скоринг текста

- Текст бьётся на предложения (по пробельным символам и знакам препинания);
- Оценка предложения – среднее геометрическое частот его n-грамм (больше-лучше);
- Если максимальную оценку получил этанол – хорошо;
- Оценка текста – процент предложений, где лидирует эталон.

Baseline

- Оценка текста по униграммной модели;
- Да, по одиночным символам;
- Зато такой Baseline легко побить!

Результаты – 1

- Замена 1 любого символа в предложении на самый частый:

n	mean percentage
1-gram	51.688 %
2-gram	89.256 %
3-gram	88.681 %

Результаты – 2

- Замена 1 любого символа в предложении на случайный из возможных:

n	mean percentage
1-gram	48.989 %
2-gram	83.741 %
3-gram	87.363 %

Что делать дальше?

- Экспериментировать с паттернами шума на корпусе и скорингом;
- Умный back-off для n-грамм при оценивании;
- Попробовать учитывать грамматические хвосты и варианты словного деления;
- Выкинуть хвосты распределения символов для оптимизации.

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
- Applying Conditional Random Fields to Japanese Morphological Analysis / T. Kudo, K. Yamamoto, Y. Matsumoto

Спасибо