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.

Текущие результаты

- Опробована сизифова методология работы, это плохая методология;
- Адаптирован корпус n-gram данных Japanese Web N-gram Version 1.

Что делать дальше?

- Ускоряться.

А конкретнее?

До	Задача
02.04	Дебаг генератора шума и скоринга
09.04	Back-off, исследование MorphoTest
16.04	Сбор статистики работы back-off
30.04	Эксперименты с опробеленными текстами
07.05	HMM, LSTM
14.05	Сбор статистики
28.05	Написание текста
...	Дописывание текста

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