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 систему исправления ошибок OCR, используя буквенную n-граммную модель языка.

Что нужно сделать?

- Выучить японский;
- Изучить имеющиеся подходы к задаче;
- Понять суть используемых методов и алгоритмов;
- Выбрать корпуса для тестирования;
- Определить Baseline;
- Улучшать качество исправления.

Что было сделано?

- Японский постепенно ботается;
- Был прочитан ряд статей по теме;
- Изучены возможности python-nltk;
- Найдены несколько корпусов (KOTOHONA, RWCP, EDR etc.);
- Сделаны прикидки для Baseline.

Текущие результаты

- Знаю хирагану и ещё чуть-чуть;
- Markov chains, accuracy metric;
- На первых порах nltk более, чем достаточно;
- Были тестовые прогоны на небольших текстах.

Что делать дальше?

- Продолжать ботать японский;
- Прогоняться на корпусах, улучшая качество;
- Экспериментировать со сложными структурами (Hidden Markov Models, Conditional Random Fields etc.);
- Возможно, перейти на C для оптимизации.

Список литературы

- Foundations of Statistical Natural Language Processing / C. D. Manning, H. Schutze.
- N-gram Language Modeling of Japanese Using Bunsetsu Boundaries / S. Chung, K. Hirose and N. Minematsu
- Applying Conditional Random Fields to Japanese Morphological Analysis / T. Kudo, K. Yamamoto, Y. Matsumoto
- Survey of Pattern Recognition Approaches in Japanese Character Recognition / S. Das, S. Banerjee

Спасибо!