Optical Character Recognition

Algorithm implementation and evaluation

Group 2
Zongbo Cai
Charlie Chen
Shiwei Hua
Xin Xia
Chao Yin

Method

- **Detection**

  Rule-based techniques

- **Correction**

  Supervised model – correction regressor

Data Cleaning

- Remove mismatch lines
- Pair tokens from tesseract with ground truth
- Remove punctuations and numbers
- Reset index

Error Detection

Rule-based techniques

rmgarbage

- more than 20 characters in length (eg. iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii...)
- punctuation characters more than alphanumeric characters (eg. ?3//la‘)
- two or more different punctuation characters in the middle of the string

(eg. b?bl@bjk.1e.322)

- three or more identical characters in a row (eg. aaaaaBlE)
- consonants more than 8 times vowels (eg. jabwqbpP)
- the first and last characters in a string are both lowercase and any other character is uppercase (eg. awwgrapHic)

New rule

- uppercase characters more than lowercase characters (eg. BBEYaYYq)
- four or more consecutive vowels / five or more consecutive consonants (eg. buauub)

Candidate Search

$$\{w_c | w_c \in \mathcal{L}, dist(w_c, w_e \leq \delta)\}$$

contains all the words in the vocabulary within a limited number of character modifications

Feature Scoring

Character-level similarity

- Levenshtein edit distance
- String similarity

Word Frequency

- Language popularity
- Lexicon existance

Contextual Concern

- Exact-context popularity
- Relaxed-context popularity

Error Correction

correction regressor

AdaBoost

- Change the underlying data distribution and classify in the re-weighted data space iteratively.
- Misclassified samples tend to receive higher weights.

Tuning Parameters

- The maximum number of estimators
- Learning rate
- Loss function

Error Correction

correction regressor

	top	precision
0	1	71.55%
1	3	86.74%
2	5	88.95%
3	10	91.85%

	typo	truth	candidate	predicted_confidence	label
141602	acrsv	acrs	acrs	0.395005	1
141598	acrsv	acrs	acr	0.077035	0
141601	acrsv	acrs	across	0.077035	0
141599	acrsv	acrs	acra	0.030645	0
141600	acrsv	acrs	acre	0.030645	0
141603	acrsv	acrs	acs	0.023050	0
141604	acrsv	acrs	acsh	0.000898	0
141608	acrsv	acrs	activ	0.000898	0
141611	acrsv	acrs	acts	0.000898	0
141679	acrsv	acrs	cars	0.000898	0

Measurement

	Measure	Tesseract	Post
0	word_wise_recall	0.645514	0.737668
1	word_wise_precision	0.651473	0.744477
2	character_wise_recall	0.921204	0.941688
3	character_wise_precision	0.950425	0.969093