Towards Whole-Book Recognition

Pingping Xiu and Henry S. Baird

Goal

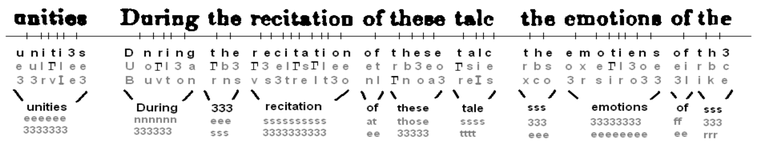
Fully automatic improvement of a recognition system for the textual contents of an entire book

Using internal evidence within a long and isogenous book, we can automatically adapt the recognizer to improve OCR accuracy.

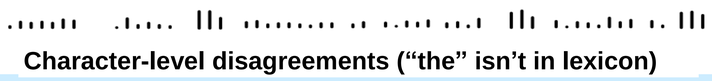
Disagreements among independent models can suggest improvements to each model

If the character classifier is inaccurate, and the dictionary is incomplete, the two models can cross-check and complement each other to improve recognition accuracy.

Quantifying Model Disagreements Using Mutual Entropy



Character recognition results and word recognition results



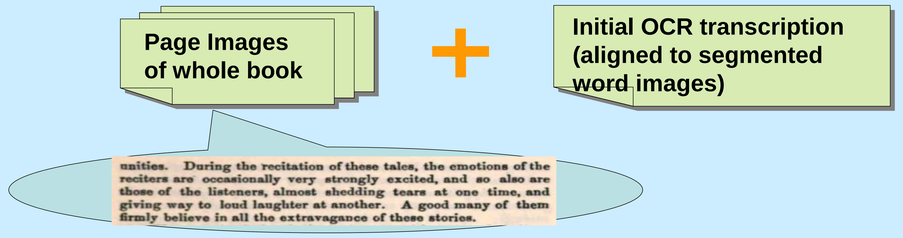
Character-level disagreements (“the” isn’t in lexicon)

Key Notion – Disagreement-Guided Model Adaptation

An information-theoretic analysis suggests that correct model adaptations will reduce the disagreement measured on the whole passage, and that the wrong ones will increase it, quite probably.

Further, we expect that the *longer* the passage is, the *more likely* this is to occur.

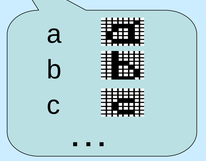
Strategy: Given a book’s images and an initial buggy OCR transcript, derive two independent models and adapt those models to that book’s images to improve the transcript.



Extract two models

Iconic model

Describes image formation and determines the behavior of a character-image classifier. Here we use template matching as our classifier.



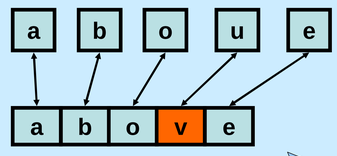
Linguistic model

Describes word-occurrence probabilities or, more simply, merely a lexicon.

Re-recognize the whole book using these two models

Suppose the iconic model suggests:

But the linguistic model prefers:



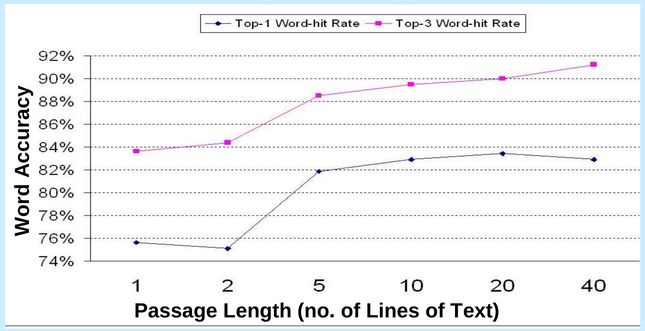
Then this disagreement suggests two possible corrections:

(1) To correct the iconic model
-- replace the label ‘u’ with ‘v’.
-- choose this if there are many suspicious words such as “uehicle”, “uictory”, “uocabulary” that have a “u” with a high disagreement.

(2) To correct the linguistic model
-- add a new entry “aboue” to linguistic model.
-- choose this if there are many other occurrences of “aboue”.

Experimental Results: A Monotonic Improvement

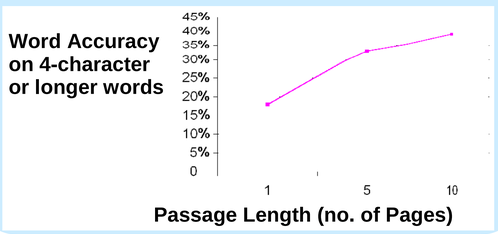
Recognition rate as a function of passage length



Tests on Much Longer Passages

An improved version of our algorithm can handle longer passages

Accuracy on 4-character and longer words, measured in a single page:



Algorithm Enhancements Described in the Paper

Using multiple templates for each character label in iconic model.

Cope with segmentation errors: one split, or one merge, per word.

Future Work

Scale up to an entire book

Compare competing policies for correcting the models

Explore branch-and-bound strategies

LEHIGH UNIVERSITY
P.C. Rossin College of Engineering and Applied Science
Computer Science and Engineering
CSE
Henry S. Baird & Daniel Lopresti
Pattern Recognition Research Lab