Enhanced OCR Scanning

Through Grammar and Word Association

Proposal

CS4440-A
Fall 2007

Group 11:
Matthew Garrison
Hung Lai
Qiming Zou

1. Motivation and Objectives

According to Wikipedia, Optical Character Recognition (OCR) “is the mechanical or electronic translation of images of handwritten or typewritten text (usually captured by a scanner) into machine-editable text.” Even though a lot of academic research has been carried out on this topic, there are still needs for improving the accuracy and completeness of the scanned text. More specifically, scanned documents are prone to inaccuracy and error. This is especially true with large amounts of aged books in libraries across the country. Our group plans on creating a solution which will help to reduce some of these errors in scanned documents.

2. Related work

a. Adaptive Post-Processing of OCR Text via Knowledge Acquisition.

Authors: Lon-Mu Liu, Yair M. Babad, Wei Sun, Ki-Kan Chan

This paper introduces new concepts in the correct identification and translation of bitmapped characters into editable text. The authors of this paper use knowledge acquisition, self learning, and adaptive ness. Their process takes in a text file generated by an OCR program and outputs a text file with fewer errors than that of the OCR text file. What sets their technique apart from others is the use of self learning. Other techniques only allow for error identification with human interaction.

b. A Filter Based Post-OCR Accuracy Boost System.

Authors: Eugene Borovikov, Ilya Zavorin, Mark Turner

The authors of this paper introduced an idea of combining several post-OCR tools to provide better translation accuracy. They introduce a system of sending the original OCR text through a series of filters. Each filter would search the text for errors and create an error log. Each error would be given a value based on the systems evaluation of the likelihood of that actually being an error. The end result is a text document with a higher percentage of correctness. They also throw in a spell checker just to give the extra correctness of the text.

c. OCR error correction using a noisy channel mode.

Authors: Okan Kolak, Philip Resnik

The approach taken by the authors of this paper is to use pattern recognition. A problem put forth by this paper was the fact that many OCR systems are un-trainable. Most have a certain set of recognizable characters and that's it. They allow their system to learn new characters. This approach not only allows for more general and flexible systems but it also allows the system to reduce error rates over time.

3. Proposed work:

A scanned document may contain some words that typical OCR post-processors cannot recognize. Our solution to this problem is to scan the document for unreadable words. If any letter is not readable, a wild card character is used as a placeholder for the letter. This word is inserted as a data point into a built in dictionary (if we cannot find an available plug-in we would construct a small dictionary for the purpose of testing our idea and for presentation purposes). This step is executed even for words with no ambiguity, since current OCR systems do not have a confidence level for each scanned word. Instead a best match is often chosen.

We will then run a clustering algorithm on the dictionary with the center being on the inserted word. The distance is then computed by a word distance metric. We then sort the cluster into an index using the distance information. Our program will also consider sentence structure and the context of the scanned word. This will be accomplished using grammar and word association rules, combined with the distance index of possible choices. The word that best fits is then chosen.

For word association rules, we plan to use machine learning and an association mining technique. The possibility of a word appearing in a sentence given other words in the sentence, and the possibility of a word appearing before/after a given word is computed. This knowledge is stored for future work, so the program's accuracy would increase with experience. We also plan to implement a currently available grammar checker to make the process more focused and efficient. We realize the algorithm could possibly be relatively slow but given the specific purpose of the program: as an extension for scanning process, it only needs to run as fast as a human can scan the document, which is not very fast at all.

4. Plan of action:

Programming language: Java (or C#), SQL (of some sort). This work is to be implemented on a Windows machine.

Proposed schedule and milestone:

Week 9:

Finalize architecture design

Week 10:

OCR implementation, word/string distance based index

Week 11:

Construction of the grammar rules, word association index system

Week 12:

Finish basic coding

Week 13:

Debugging and testing

Week 14:

Finishing debugging and testing

Week 15:

Final tune up

5. Evaluation and Testing Method

Black box testing:

The resources from Georgia Tech’s library shall be used to experiment and evaluate the actual process of scanned document.

Evaluation:

We will simply compare the accuracy with standard OCR implementations such as Adobe Acrobat and Microsoft Image Reader.

Bibliography

1) Lon-Mu Liu, Yair M. Babad, Wei Sun, Ki-Kan Chan. “Adaptive Post-Processing of OCR Text via Knowledge Acquisition.” ACM Annual Computer Science Conference Proceedings of the 19th annual conference on Computer Science. Pages: 558 – 569. 1991

2) Eugene Borovikov, Ilya Zavorin, Mark Turner. “A Filter Based Post-OCR Accuracy Boost System.” Conference on Information and Knowledge Management Proceedings of the 1st ACM workshop on Hardcopy document processing Pages: 23 – 28. 2004

3) Okan Kolak, Philip Resnik. “OCR error correction using a noisy channel mode.” Human Language Technology Conference Proceedings of the second international conference on Human Language Technology Research. Pages: 257 – 262. 2002

4) Wikipedia. “Optical Character Recognition.” September 29, 2007 http://en.wikipedia.org/wiki/Optical_character_recognition