NATIONAL LIBRARY OF AUSTRALIA

Increasing the Accuracy of OCR

Author: Rose Holley (Manager - ANDP)
Version: 1.0
Date: 20 August 2008

Introduction

Since obtaining the first OCR files back from the OCR Contractor the Library has been conducting further research into how OCR accuracy can be improved. The OCR accuracy from newspaper pages can be relatively low due to the poor quality of the original source (microfilm and papers).

The Library has been particularly interested in finding out if improved OCR accuracy can be obtained from greyscale rather than bi-tonal tiff images. To this end the Library has performed tests internally and tests with 2 scanning/OCR contractors.

Test Method

In each case the same files were used and the same version of OCR software (Abby Finereader 8). 45 pages were used for testing (a bi-tonal and a greyscale of each). The Library did not use pages from a single title on a single date for the tests, instead using a representative sample of good, average and poor quality pages from different newspapers and different dates (as it really would be in production mode). It is worth noting that on a single newspaper title from a single date there may be a difference in results.

Interestingly although contractors at the start of the testing had indicated that they could process greyscale files and that these would yield better results, in actuality this was not the case. Both the contractors converted the greyscale files supplied into image optimised bi-tonal files for the OCR process, rather than using the actual greyscale file. This was because greyscale files take much longer to process, handle and return due to their large size. If greyscale files were processed day to day the cost would be much greater than that of bi-tonal files. For research purposes the Library still wanted to find out if OCR accuracy would be significantly better with greyscale files. The Library therefore asked for the actual greyscale files to be processed (rather than converted), and also for the greyscale files to be converted into image optimised bi-tonal files and processed. The Library would then also be able to compare our own image optimised bi-tonal files with those created in the tests. The image optimised bi-tonal file in the tests should be created by running files through a generic image optimisation program as it would be in real production, rather than handcrafting the example for perfection.

Results

The results showed there was no *significant improvement* in OCR accuracy between using greyscale or bi-tonal files. In addition the image optimised bi-tonal files created by the Contractors and the Library tests did not result in *significantly* improved OCR compared to those created by the Library's scanning contractor and normally used. (The Library's scanning contractor creates the bi-tonal files as part of the scanning process using NextStar software). The overall average OCR accuracy results of the sample were as follows:

Good quality pages	98% character accuracy raw OCR text
Average quality pages	93% character accuracy raw OCR text
Poor quality pages	71% character accuracy raw OCR text

The average greyscale pages were examined more closely on an individual basis and the following information was obtained:

There was 0.0 -3.0 % character accuracy variancy compared to bi-tonal files. A single file had a -3.0% variancy compared to bi-tonal. This meant that sometimes the greyscale yielded a slightly better result than bi-tonal, sometimes no different, and in one case worse.

Overall the bi-tonal average OCR accuracy score for the average sample was 93.8% and the greyscale average score was 94%.

This small variation would not warrant the extra cost involved in processing greyscale files and would not lead to uniformly improved OCR accuracy rates on every file.

Outcomes
It was agreed by everyone taking part in the tests that only 3 things could improve the accuracy of the OCR and these were:

1. Improvement in quality of original source
2. Making manual adjustments to image optimisation for each file
3. Manual intervention in the OCR process for each file to improve results

The first item was not possible to improve or change and the second and third items are not cost effective or efficient when processing millions of pages with Contractors.

At this stage the Library discussed again whether the public should be able to view the original OCR text in the search system and what purpose it would serve, especially if a lot of it was very poor quality. A member of the team then suggested that instead of manual intervention by the Contractors to improve OCR accuracy why not manual intervention by the public users of the service? For example a user could view, edit, correct and save OCR text on whatever article they happened to be looking at. This was an exciting and groundbreaking idea and to the best of our knowledge it had not been implemented by any other newspaper service. The major benefits of allowing the public users to manually correct OCR text would be:

1. Text could be improved and potentially whole articles made perfect, therefore improving the searching.
2. It would not cost the Library any additional money in contractor processing since the public would do it for free.
3. The community could be harnessed to improve and add value to the service in a way we had not previously thought of (in a similar way to Wikipedia).

Things we would need to consider would be:

1. How hard technically would this be to implement?
2. Would the public really do it?
3. Should/could we moderate it?

The team decided to action the idea. OCR correction by users was implemented and tested in the prototype search system released to State and Territory Libraries for testing in December 2007. It was positively received, though most Libraries queried if and how moderation may take place. It was then implemented in the Beta search system which was released to the public on 25 July 2008. In the first month of use the following information has been noted:

1. Users did not expect to be able to correct OCR text and it was initially difficult to convey the concept, purpose and technique of doing this.
2. Once users discovered OCR correction they found it addictive and rewarding and were actively correcting much more than we had expected.
3. OCR correctors want to correct the odd word here and there as they read articles, and/or correct the entire article.

4. Half of the corrections were done by anonymous users (i.e. only some OCR correctors decided to login and register themselves).
5. In the first month of use over 200,000 lines of text have been corrected by the public.
6. OCR correction is not being moderated by the Library at this stage.
7. The goodwill of people and their willingness to work on a worthy cause is quite amazing.
8. The quality of the resource is being greatly improved as users contribute in this way.

Summary

Carrying out these OCR tests has led the Library to a greater understanding of the OCR process and options for improving OCR accuracy. The Library had initially been working on the assumption that improving the OCR accuracy relied on OCR contractors processes. Once the idea of public users correcting OCR text to improve it was established it took off. The Library now consider this the most efficient and cost effective way to improve OCR accuracy, and it also has the big advantage of building an involved and empowered user community. The Library will be seeking more ways to interact with the community and to further develop and improve the technique for OCR correction within the service over the coming months.