UNIVERSITY POLITEHNICA of BUCHAREST
DEPARTMENT OF COMPUTER SCIENCE

Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs

"Many Hands Make Light Work"

Lupoae Mihai

lu_mi_nos@yahoo.com

Cuprins

- Australian Newspapers Digitisation Program
- Functionarea OCR

- Factori care influenteaza OCR
- Masurarea acuratetii OCR
- Metode de imbunatatire a OCR
- Teste si rezultate
- Corectarea manuala a textului OCR

Australian Newspapers Digitisation Program

- Este un program al Librariei Nationale Australiene. Scopul acestui program este digitizarea ziarelor australiene la scara mare si imbunatatirea acuratetii OCR. Include ziare din perioada 1803-1954.
- Sunt prezentate modul in care functioneaza OCR pentru ziare, factori care influenteaza acuratetea OCR, metode de masurare si imbunatatire a acuratetii si solutii viabile pentru proiecte de digitizare la scara mare.

Functionarea OCR

- Oamenii pot recunoaste text cu fonturi si calitati de imprimare diverse si pot aplica abilitatile lingvistice si cognitive pentru a traduce textul in cuvinte. Poate scana layout-ul, sectiunile, titlurile, si poate citi articolele in ordinea corecta.
- Programele OCR pot face aceste lucruri, dar nu la fel de bine ca un om. In plus are nevoie de contrastul alb-negru pentru a distinge textul de fundal.

Functionarea OCR

- Pasii OCR
  - Preprocesare: de-skewing, de-speckling, etc.
  - Analiza structurii paginii: imparte pagina in elemente(coloane de text, tabele, imagini). Liniile sunt impartite in cuvinte, apoi in caractere, care sunt comparate cu niste imagini din baza de date si primesc o nota de incredere. (mai mult)
  - In plus se poate face o analiza a cuvintelor si nota de incredere poate creste. (mai mult)

Factori care influenteaza OCR

- Obtinerea sursei
  - calitatea sursei (mai mult)
- Scanarea paginii
  - Rezolutia si formatul fisierului (mai mult)
- Preprocesarea imaginii(creare de contrast alb-negru)
  - Bit-depth
  - Binarizare
  - Calitatea sursei (mai mult)
- OCR-Analiza layout-ului
  - Pagini oblice
  - Pagini cu layout complex
  - Spatii albe intre linii, coloane. (mai mult)
- OCR-Analiza marginii fiecarui caracter
  - Optimizarea imaginii
  - Calitatea sursei (mai mult)
- OCR-Potrivirea caracterelor cu imaginile din baza de date.
  - Imaginile din baza de date
  - Algoritmii OCR
- OCR-Potrivitea cuvintelor cu dictionarul.
  - Algoritmi si dictionarul OCR

Masurarea acuratetii OCR

- Software-ul OCR calculeaza un nivel de incredere pentru fiecare caracter detectat.
- Increderea pentru cuvant si pagina poate fi calculata folosind algoritmi din software-ul OCR sau externi.
- Doar un om poate sti daca un caracter este corect.
- Increderea si acuratetea sunt diferite, dar in practica acuratetea e inlocuita de incredere deoarece este convenabil pentru cantitati mari de text.

Masurarea acuratetii OCR

Metoda de obtinere a acuratetii pentru articole:

- Stabileste acuratetea pentru o mostra OCR manual.
- Obtine increderea pentru acea mostra.
- Scrie un algoritm pentru a le corela.
- Foloseste algoritmul pentru a obtine acuratetea pentru articol.
- Verifica algoritmul la intervale regulate.
- Good average poor acurracy

Metode de imbunatatire a OCR

- Imbunatatirea calitatii sursei
- Scanare la 300 dpi sau mai mult
- Folosire fisiere tiff
- Ajustari manuale pentru procesul optimizare pentru fiecare pagina
- Folosirea celui mai bun software de optimizare OCR
- Experimentare cu fisiere greyscale in loc de alb-negru
- De-skew
- Interventie manuala in procesul OCR pentru fiecare fisier
- Antrenare OCR
- Votare: folosirea mai multor solutii OCR si alegerea celor mai bune rezultate
- Folosirea dictionarelor australiene pentru procesul OCR
- Corecteaza textul OCR manual.
- Folosirea matricei de confuzie la procesarea OCR.

Teste si rezultate

- Pentru testare au fost folosite 45 de pagini de ziare din perioada 1803-1954.

- Folosirea dictionarului australian

A dat rezultate mai proaste decat folosirea exclusiva a dictionarului incorporat. (mai mult)

- Folosirea fisierelor greyscale
  - 93.8% alb-negru, 94% greyscale
  - Diferenta este prea mica si prea inconsecventa pentru a folosi o metoda mai costisitoare pentru greyscale.
- Compararea software-urilor de optimizare
  - Diferenta intre software-urile testate este neglijabila.
- Matricea de confuzie
  - Erorile OCR(ex: Sydnoy in loc de Sydney) sunt modelate intr-o matrice de confuzie care poate fi folosita pentru a imbunatati OCR. (mai mult)
- Corectarea manuala a textului OCR
  - Aceasta corectare este facuta de utilizatorii acestui serviciu.

Corectarea manuala a textului OCR

- Beneficii
  - Imbunatatirea calitatii
  - Comunitatea este implicata in imbunatatirea resursei
  - Este ieftin
  - Sunt construite noi comunitati virtuale

Corectarea manuala a textului OCR

- Riscuri
  - Aceasta functionalitate nu a mai fost implementata si nu au fost stabilite reguli
  - Vandalism
  - Multe corectari pot compromite baza de date
  - Utilizatorii nu fac corectari si timpul de dezvoltare este irosit
  - Utilizatorii nu inteleg conceptul de corectare a textului
  - Utilizatorii sunt derutati de diferentele intre corectarea textului, comentarii si tag-uri.
  - Dificultatea si durata implementarii
  - Interfata potrivita pentru utilizatori (mai mult)

Corectarea manuala a textului OCR

- Implementare
  - Textul OCR este furnizat intr-un fisier ALTO XML. Informatia este pastrata intr-o baza de date SQL
  - Prima parte din articol este corectata de furnizor. Informatia este pastrata in baza de date SQL
  - Corectiile facute de public sunt pastrate in baza de date.
  - Informatia este pastrata in format binar pentru a minimiza spatiul si timpul de parsare.
  - Pentru fiecare pozitie a unui cuvant sunt pastrate
    - Coordonatele
    - Cuvantul original OCR
    - Cuvantul corectat de furnizor
    - Cuvinte cu cratima reconstruite
    - Corectari ale publicului
  - Motorul de cautare este Lucene http://lucene.apache.org/java/docs/index.html
  - Versiunea publica a articolului este (in ordine):Ultima corectare, textul corectat de furnizor, textul OCR.

[In order to guard against Imposition, notices of
L Birth», Marriages, and Deaths muat bo authenti

cated by somo rcspcctablo person In Melbourne, to
secure their insertion]

Binns.

Adambon -On the 2Dlh ult, at the Bank o [\ letona,

Menno, the wife of J A. Adamson of a daughter

Gladstones -On the 2nd Inst, at Kew, Mrs. John

Gladstones of a son

Harris -On the 30th ult, at \o 2 Victoria buddings,

Queen street, the w ito of Abraham Harris of a sou

RICHARDSON. -On the 1st inst., at Preston-house,

PrcBton, tho wife of Mannaduko N

secure their insertion.]

Births.

ADAMSON.—On the 29th ult., at the Bank of Victoria, Merino, the wife of J. A. Adamson of a daughter.

GLADSTONES.—On the 2nd inst., at Kew, Mrs. John Gladstones of a son.

HARRIS.—On the 30th ult., at No. 2 Victoria-buildings, Queen-street, the wife of Abraham Harris of a son.

RICHARDSON.—On the 1st inst., at Preston-house, Preston, the wife of Marmaduke N. Richardson, late Lieut. H.M. 83rd Regt., of a son.

Marriages.

GEORGE—SISELY.—On the 1st inst., by special licence, by the Rev. S. Williams, Wm. Jno. George, teacher, Wilmington, S.A., to Kate, only daughter of Mr. James Sisely, of Kangaroo-flat.

MITCHELL—WRIGHT.—On the 24th ult., at the residence of the bride's parents, by the Rev. Thos. Jones, Thomas Mitchell, of Sydney, to Elizabeth Anne, only daughter of Wm. Wright, of Queensberry-street, Carlton. Sydney papers please copy.

M'CLELLAND—SPENCER.—On the 2nd inst., at Carlton-street, Carlton, by the Rev. Alexander Yule, Samuel M'Clelland, of Lygon-street, Carlton, to Ellen Spencer, of Latrobe-street, Melbourne.

O'LEARY—WINSTANLEY.—On the 29th ult., at 16 Milton-street, West Melbourne, by the Rev. E. T. Miles, Daniel O'Leary, of Williamstown, to Mary Ellinor, second daughter of John R. Winstanley, late of Adelaide, South Australia.

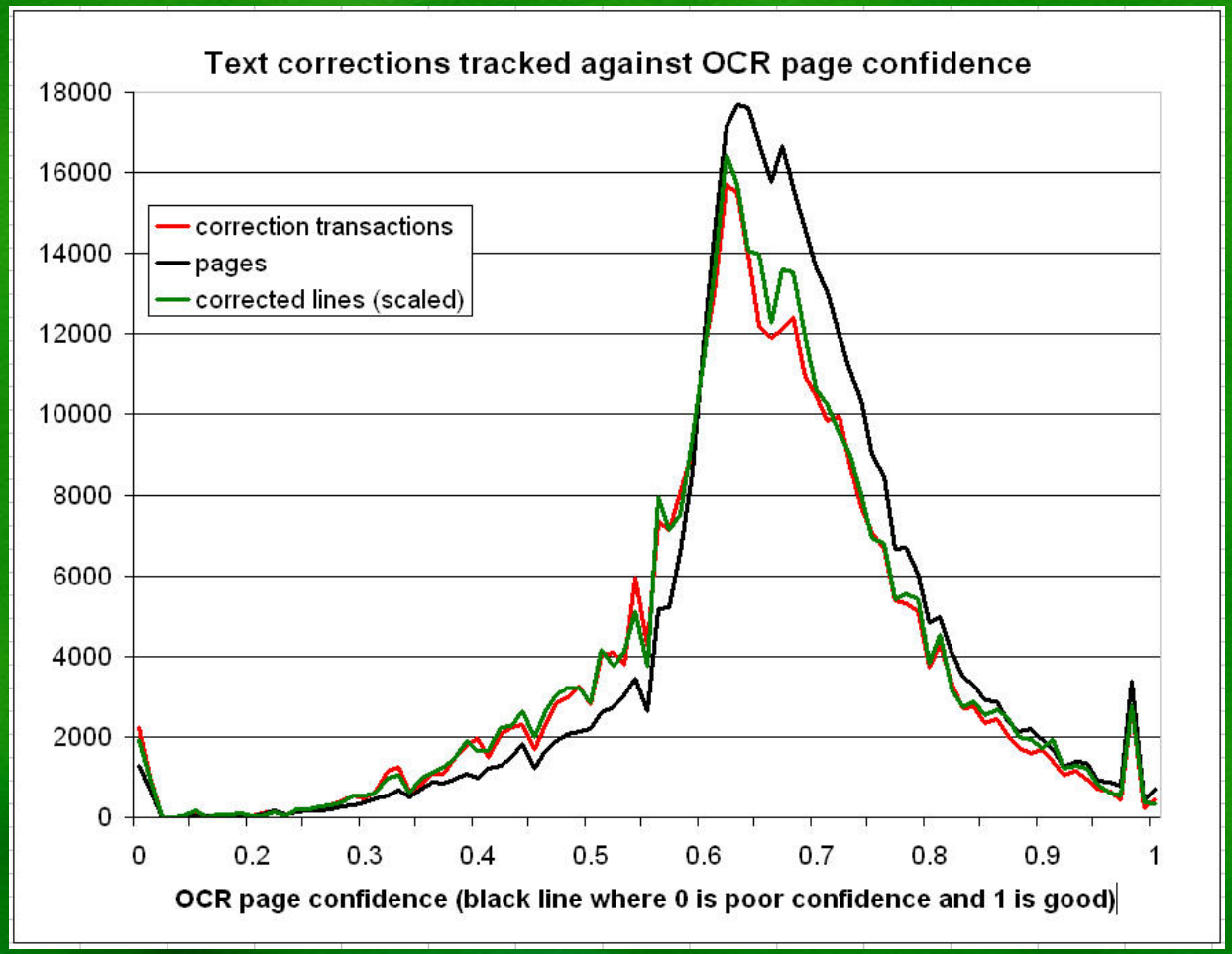
Deaths.

ASTLEY.—On the 2nd inst., at Rowena-street, Richmond, Emily, youngest daughter of William and Mary Ann Astley, aged eight years.

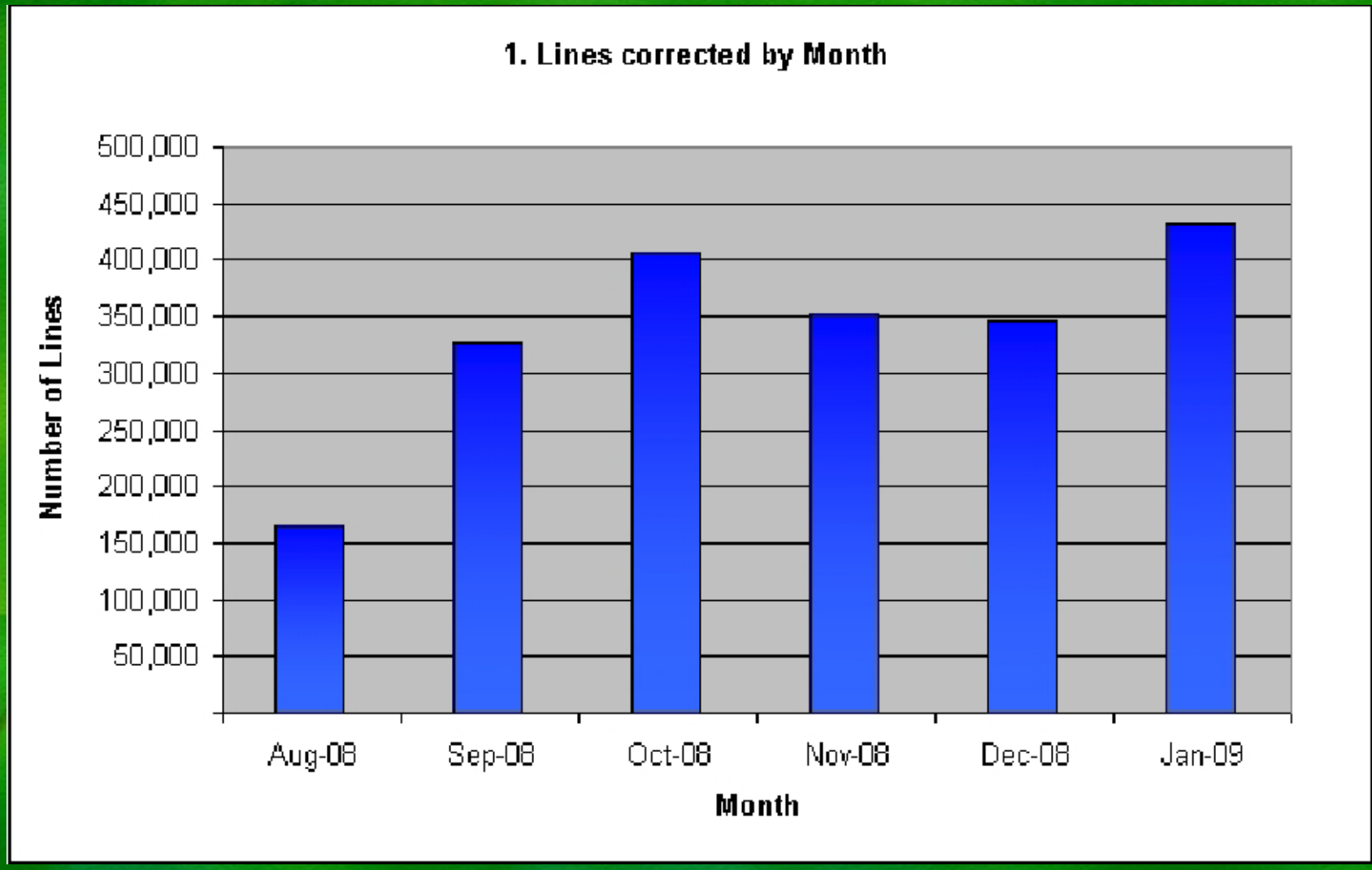
BEHAN.—On the 11th ult., at Tangihanga, near Gisborne, New Zealand, William Behan, aged 70 years, late of Queenscliff and Geelong.

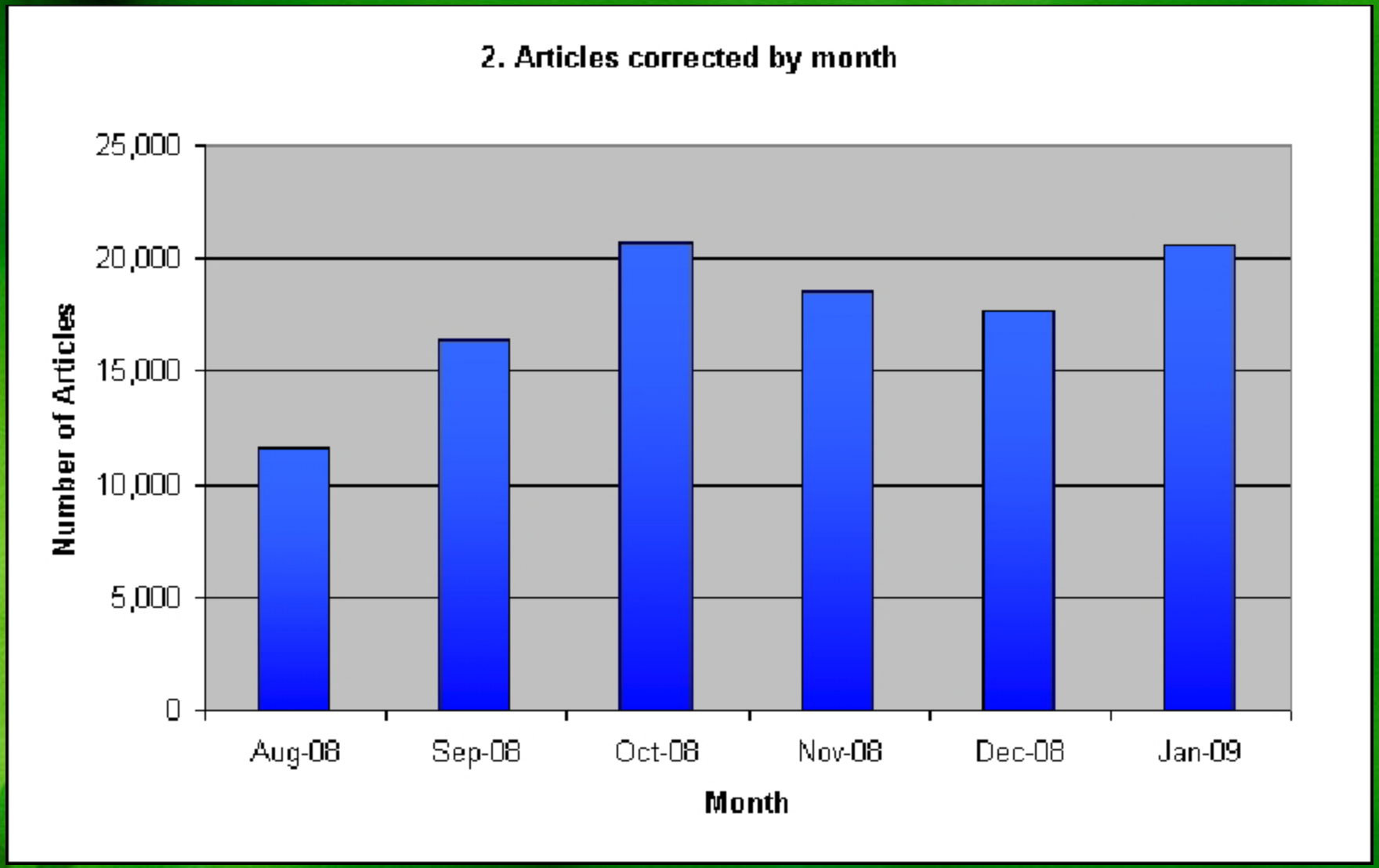
COOK.—On the 8th November, at Islington, London, Ann Cook, aged 80 years, the dearly-beloved mother of William Cook, pawnbroker, Beach-street, Sand-

Masurarea imbunatatirii acuratetii articolelor



Niste tabele




2. Articles corrected by month
25,000
20,000
15,000
10,000
5,000
0
Number of Articles
Aug-08
Sep-08
Oct-08
Nov-08
Dec-08
Jan-09
Month

	At Nov 2008	At Feb 2009
Number of registered users	1488	2994
Lines of text corrected	1 million	2.2 million
Number of articles corrected	60,000	104,000
Top corrector	60,000 lines in 2000 articles	103,000 lines in 2600 articles
Number of comments added	800	1806
Number of tags added	18,000	43,000
Unique visitors to site	94,000	205,000
% of users in Australia	78%	75%
Content in service	367,000 pages 3.5 million articles	367,000 pages 3.5 million articles

Concluzii

- Principalul motiv pentru acest proiect a fost imbunatatirea calitatii datelor.
- Pentru utilizatori comunitatea virtuala si puterea de a contribui sunt aproape la fel de importante ca datele de calitate mare.
- Utilizatorii au demonstrat dorinta de a lucra pentru un bine comun. Aceasta comunitate este mai performanta decat Libraria Nationala Australiana.
- Latest news: Din octombrie programul are peste 30 000 000 articole
- 9000+ membrii au corectat 12.5 milioane de linii din ziare.


What's up, Doc ?

The text correction user interface operates as follows:

1. When a line of text is right clicked or the "help fix this text" mouse over is clicked, JavaScript in the
browser puts the text of the line into an input text box. The user can overtype/edit the text.
2. When the user hits ‘enter’ on the keyboard the cursor is moved to the next line and a new input
box is created.
3. When ‘save’ is clicked, JavaScript on the client creates a list of changed lines and sends them to the
server, using AJAX (JavaScript issuing a HTTP request). Each changed line is preceded with the
XY co-ordinates of the 1st word in the line which lets the server match updated lines.
4. The server code matches up the existing and changed lines. For each changed line, it tries to ‘line
up’ the old and new words so that the new words can be highlighted on the page image - that is,
the system uses the bounding box of old words to correctly highlight new words. This is a hard
problem in general which we solve heuristically by matching the longest runs of matching
characters. But it is far from perfect, and doesn’t work very well when the number of words in a
line is significantly changed (words inserted/combined/deleted).
5. The server does not allow lines to be inserted or deleted. (inapoi)

Implementation of dictionaries seemed to be a potentially quick and easy way to increase OCR accuracy, which is why this solution was tested first. However, we quickly realised that using the Macquarie Australian national dictionary8 (first published in 1981) was unlikely to make any difference to OCR results, since at the time Australian newspapers were published (1803-1954) pure English was being spoken primarily, and there were few Australian colloquialisms that might appear in those newspapers. Jumbuck (sheep), Billabong (watering hole) and tucker (food) were the only possible exceptions we could think of. However, aboriginal place names were in wide use then, as they are today, so we asked our OCR contractor to incorporate the official Australian gazetteer of place names as a secondary dictionary into ABBYY (the primary dictionaries are already built in) and run 45 sample pages through OCR. We then compared the results with pages using the primary dictionary only and pages using no dictionaries.

Our findings were that using ABBYY primary dictionaries gave the highest accuracy, followed by using no dictionaries. Surprisingly, the worse results were obtained from using the primary and secondary dictionaries together. This was unexpected and at variance with experiences from the National Library of New Zealand, which had obtained improved results when applying a secondary dictionary populated with geographic place names. We wanted an explanation for this, but unfortunately, both our contractor and we had a very limited understanding of when and how a secondary dictionary was applied and utilised. The OCR contractor approached ABBYY about this, but ABBYY was reticent to share that proprietary software information. We suspected that the secondary dictionary was interfering with the primary dictionary or had been implemented incorrectly, but we were unable to obtain any further information about implementation of dictionaries or application of secondary dictionaries, so the results could not be explained. We repeated the tests again on different pages six months later with the same contractor but the results were the same. More time and research on the use of two dictionaries is desirable, because there were a lot of unknowns for us in the test process. In the meantime, however, we have decided to use the default primary dictionaries only in ABBYY for OCR of Australian Newspapers. (inapoi)

OCR errors, whilst not predictable, can be modelled statistically and are very different from human spelling mistakes. A confusion matrix would model these errors in order to be able to correct them and improve OCR. For example, an OCR engine may commonly mistake the letter h for l i or m for in. Thus the word 'the' would be translated as 'tlie' instead of 'the'. In the ANDP the mistranslated word 'tlie' occurs in 1 of 8 articles. But the mistranslations are more significant on words on which users may search. There are 30,000 articles with 'Sydnoy' instead of 'Sydney'. These observations of unigrams, bigrams (pairs of characters) and trigrams and what they get translated as form the basis of the confusion matrix. The matrix could be applied across the OCR as the first part of the process by identifying correction candidates.

Once a word is formed, its occurrence in a sentence in the language model can be applied. Words are more likely to occur in context; therefore, a sentence reading "the cot sat on the met" is much more likely to actually be "the cat sat on the mat". Likewise a "rich coal scam" is more likely to be a "rich coal seam", the occurrence of the word scam directly next to coal is unlikely. The confusion matrix applied with a language model has the potential to increase OCR accuracy, though not to make it perfect. The team were interested in this approach but did not have the time, resources or adequate OCR data to pursue it further. Software development would be required and thorough testing of data using the matrix and algorithms to ensure that results were improved not worsened. It was unclear how long the post OCR processing might take and how viable it may be for large scale newspaper projects. In 2008 the Impact12 Project was established in Europe as part of the i2010 vision of the European Digital Library. It has funding of 15.5 million Euros and has 15 national library partners. Some of the aims of the project are to develop text recognition products further, develop post correction modules and investigate the most effective methods of enhancement and enrichment of OCR for mass scale digitisation projects. The confusion matrix and post language processing may fall under this research. (inapoi)

It analyzes the stroke edge, the line of discontinuity between the text characters, and the background. Allowing for irregularities of printed ink on paper, each algorithm averages the light and dark along the side of a stroke, and advances numerous hypotheses about what this character is. Finally, the software makes a best guess decision on the character. This character is given a confidence rating. The encoding of this confidence is dependent on the software or schema used to represent the OCR results. Therefore, a confidence rating encoded according to the ALTO standard6 for newspapers is an integer within the range of 0-9, 9 being very confident. (inapoi)

The question of what is acceptable has not been answered, but in speaking to other libraries and OCR contractors, it was generally agreed for historic newspapers that when we talk about good, average and bad OCR we mean:

Good OCR accuracy = 98-99% accurate (1-2% of OCR incorrect)

Average OCR accuracy = 90-98% accurate (2-10% of OCR incorrect)

Poor OCR accuracy= below 90% accurate (more than 10% of OCR incorrect)

However, there was not consensus on whether these percentages referred to character or word confidences, and whether this was at page or article level, and many people were still confused about what accuracy/confidence meant and how it was calculated. We concluded that we could not set a baseline figure for acceptable accuracy until the method for measuring it had been firmly established and agreed. (inapoi)

The built-in English dictionaries and possibly dictionaries of other languages are checked to see if the word matches. If it does, the confidence rating of the characters may be increased. The built-in dictionaries have a complicated relationship with the algorithms and the hypotheses, and how they integrate together is usually kept confidential by the software companies. (inapoi)

Obtain original source

- Use original hard copies if budget allows (digitisation costs will be considerably higher than for using microfilm).
- Hard copies used for microfilming/digitisation should be the most complete and cleanest version possible – preferably not bound.
- Use microfilm created after establishment and use of microfilm imaging standards (1990's or later).
- Use master negative microfilm only (first generation) or original copies, no second generation copies. (inapoi)

Scan file

- Rezolutia de scanare ar trebui sa fie mai mare de 300 dpi
- Formatul fisierului este lossless(tiff). (inapoi)

Create good contrast

- Scaneaza imaginile greyscale sau alb-negru
- Optimizare a imaginii inaintea procesului OCR. (inapoi)

OCR software-Layout of page analysed and broken down.

- De-skew.
- Layout-ul paginilor nu poate fi schimbat. Ghinion. (inapoi)

OCR software - Analysing stroke edge of each character

- Marginile caracterelor trebuie “smoothed, rounded, sharpened, contrast increased” inainte de OCR (inapoi)

Masurare (inapoi)

When pages are supplied to the Library from the OCR contractor the accuracy of OCR text is not being measured due to difficulties with being able to do this 9. However the page level OCR engine 'word confidence' figure is provided in the ALTO file. It is unknown how closely this OCR engine word confidence level may match the real word accuracy. The figure provided for word confidence is between 0 and 1, where 0 is considered very low confidence and 1 very high confidence. Of the

---

Many Hands Make Light Work, Rose Holley, March 2009 Page 24 of 28

360,000 pages currently in the Australian Newspapers service the page confidence levels are as follows:

7.1% of pages have a confidence of < 0.5
18.5% of pages have a confidence of < 0.6
61.0% of pages have a confidence of < 0.7
87.1% of pages have a confidence of <0.8

Therefore 42.5% of pages have an "average" confidence between 0.6 and 0.7

The ANDP team would like to be able to measure the improvement in overall accuracy of articles or the corpus as a whole now that public text correction is taking place. However this is still not possible to do due to lack of resource. It could be done simply by comparing words in an article with words in a dictionary before and then after text correction as a comparison.

In the meantime, as a matter of interest, all of the OCR text corrections in articles on a page (by line and by transaction i.e. clicking save) have been plotted against the existing OCR engine provided page confidence levels, for the entire 360,000 pages. We wanted to see if the lower the confidence the higher the correction transactions. The corrected lines have been scaled back by a factor of 7 so that they are more easily compared. The graph shows that corrections are "above" the page number curve for low confidence, and "below" the page number curve for high confidence, and about the same for midconfidence pages (between about 0.6). So, lower confidence pages tend to attract slightly more corrections proportionally than higher confidence pages, but the effect isn't that pronounced. Pages with very low confidence of between 0.3 and 0.55 make up 10.6% of the corpus and they get 16.7% of the corrections, pages with high confidence between 0.75 and 1 make up 20.4% of the corpus and they get just 18.4% of the corrections. 69% of the corpus is of average confidence between 0.55 and 0.75 and these pages get 64% of the corrections.

It would be entirely feasible as some users have suggested to actively "dish up" the articles on pages that have a low page confidence if we wanted to target these for correction.

Risk	Mitigation/Solution
No such functionality has been implemented by a Library before and there are no policies in place	• Develop a disclaimer and terms of use for the service • Test in a beta version
Potential vandalism of text	• User can always see original image of page so will not unknowingly view vandalised text • Roll back to original OCR raw version on identified cases/entire database • Disable correction functionality if vandalism occurs • Make login mandatory instead of optional and then block specific users • Rely on the public to either not vandalise or to report vandalism • Develop a moderation module if required once adequate testing has taken place in beta
Large amounts of text correction activity compromise database/service	• Make beta a soft launch, no press releases, just word of mouth so that usage increases gradually
Users don't do text correction and development time is wasted	• Test in beta for several months • Implement basic correction only without a moderation module until it is proven that users want to be involved. Do further development after proven. • Provide information about how it will help everyone else if users participate • Encourage contributor libraries to promote it to their users • Promote service
Users don't understand the concept of text correction	• Give users adequate time to understand the concept and make beta available for several months. • Don't use the term 'OCR' ; instead use a term that was understood in user testing 'electronically created text' • Liken concept to that of Wikipedia • Improve interface and add help
Users are put off using the service because of seeing all the raw OCR needing correction which may be 'gibberish' and clutter screen	• Put a frame splitter in beta so users can hide or show raw OCR and change the width of it on screen. Therefore some users can view newspaper page image only and some the full screen of OCR text.
Users are confused about the difference between text correction, adding comments and adding tags and how to do it	• User testing before release of beta • Terms tested on users • Make functions clearer on interface • Add 'what is this' on interface • Add help text • Populate some articles with examples of the features before releasing Beta
Not sure how technically difficult it will be to implement text correction and how long it will take	• Allocate 2 weeks for development work and see what can be achieved in this time.
Not sure what an effective user interface for text correction will be like and how it can be developed since this has never been done before.	• Use expert user interface designer • Test interface with public users