Keep, Change or Delete? Setting up a Low Resource OCR Post-correction Framework for a Digitized Old Finnish Newspaper Collection

Kimmo Kettunen

Center for Preservation and Digitisation, National Library of Finland, Mikkeli, Finland
kimmo.kettunen@helsinki.fi

Abstract. There has been a huge interest in digitization of both hand-written and printed historical material in the last 10–15 years and most probably this interest will only increase in the ongoing Digital Humanities era. As a result of the interest we have lots of digital historical document collections available and will have more of them in the future.

The National Library of Finland has digitized a large proportion of the historical newspapers published in Finland between 1771 and 1910 [1,2,3]; the collection, Digi, can be reached at http://digi.kansalliskirjasto.fi/. This collection contains approximately 1.95 million pages in Finnish and Swedish, the Finnish part being about 837 million words [4]. In the output of the Optical Character Recognition (OCR) process, errors are common especially when the texts are printed in the Gothic (Fraktur, blackletter) typeface. The errors lower the usability of the corpus both from the point of view of human users as well as considering possible elaborated text mining applications. Automatic spell checking and correction of the data is also difficult due to the historical spelling variants and low OCR quality level of the material.

This paper discusses the overall situation of the intended post-correction of the Digi content and evaluation of the correction. We shall present results of our post-correction trials, and discuss some aspects of methodology of evaluation. These are the first reported evaluation results of post-correction of the data and the experiences will be used in planning of the post-correction of the whole material.

Keywords: historical newspaper collections, OCR post-correction, evaluation

adfa, p. 1, 2011.

1 Introduction

Newspapers of the 19th and early 20th century were many times printed in the Gothic (Fraktur, blackletter) typeface in Europe. It is well known that the typeface is difficult to recognize for OCR software [5, 6]. Other aspects that affect the quality of the OCR recognition are the following, among others [6, 7]: quality of the original source and microfilm, scanning resolution and file format, layout of the page, OCR engine training, etc.

As a result of these difficulties scanned and OCRed document collections have a varying number of errors in their content. The number of errors depends heavily on the period and printing form of the original data. Older newspapers and magazines are more difficult for OCR; newspapers from the 20th century are easier (cf. for example data of [8] that consists of a 200 year period of The Times of London from 1785 to 1985). There is no clear measure of the number of errors that makes the material useful or less useful for some purpose, and the use purposes of the digitized material vary hugely. A linguist who is interested in the forms of the words needs as error free data as possible; a historian who interprets the texts on a more general level may be satisfied with text data that has more errors.

OCR errors in digitized text collections may have several harmful effects, one of the most important being possibly worse searchability of the documents in the collections. Ranking of the documents in search result list is usually clearly harmed. With high-enough word level accuracy of the OCRed collections searchability is not harmed significantly according to Taghva et al. [9]. Tanner et al. [10] suggest that word accuracy rates less than 80 % are harmful for search, but when the word accuracy is over 80 %, fuzzy search capabilities of search engines should manage the problems caused by word errors.

Other effects of poor OCR quality will show in the more detailed processing of the documents, such as sentence boundary detection, tokenization and part-of-speech-tagging, which are important in higher-level natural language processing tasks [11]. Part of the problems may be local, but part will cumulate in the whole pipeline of NLP processing causing errors. Thus the quality of the OCRed texts is the cornerstone for any kind of further usage of the material.

2 Framework of Post-correction

Correction of the OCR output can be done interactively during the OCR process and interactively or non-interactively after the OCR process has finished. Then it can be based on crowdsourcing or automatic correction. Crowdsourcing with the Digi data has been tried with small amount of data [12], but as the amount of data is over 800 million words, this approach is clearly not feasible. It is obvious that partial or total re-OCRing and automatic post-correction are the only realistic ways of improving the quality of the data. We shall concentrate on the automatic post-correction in our discussion.

In [4] we evaluated the quality of the Finnish Digi data with seven smallish samples which have about 212 000 words altogether. The results of the evaluation show that the quality varies from about 60 % word accuracy at worst to about 90 % accuracy at best. The evaluation samples, however, are small, and it is hard to estimate what the overall quality of the data is. We expect it to be somewhere in the range of 50–80 % accuracy, but there may be a lot of variation. As the spelling error examples of [4] show, there are lots of really hard misspellings in the data, up to Levenshtein distance of 8 and even further from that. Morphological analysis with a state-of-the-art morphological analyser showed that only 4.3 % of the word types in the collection are recognized by the analyser. The same analyser recognizes 58.1 % of the word types from the same period in a good quality hand edited collection, and so it is obvious that a substantial part of the unrecognized words are misspellings, not out-of-vocabulary words. A high percentage of hapax legomena words (73.4 %) among the unrecognized words substantiates this, too.

2.1 Post-correction

Our discussion of OCR post-correction in this paper concerns non-word error detection and isolated word error correction as defined in Kukich [13]. Non-word detection detects words that do not occur in the used dictionary or wordlist. Isolated word correction tries to correct single words out of their context. There are several techniques for doing this, and Kukich [13], for example, lists six different approaches. In our result examples we will show results of one particular technique, minimum edit distance aka Levenshtein distance. In this phase we do not

aim to do real-world spelling-correction, i.e. context sensitive word correction, as this would clearly be out of the scope of our means and resources.

OCR result evaluation and post-correction evaluation are based on character level match between the characters of the output of the OCR results and the original "error free" data. The originals used as the comparison – many times known as ground truth - are usually hand-edited material or good quality parallel digital versions of the material. Due to lack of availability of high quality comparison material, evaluations of the digitation process and its post-correction are mainly based on quite small samples, which is inevitable.

2.2 Evaluation Data

As evaluation data we use six of the seven parallel samples used in [4]. One of the samples is too large to be used with the OCR Frontiers toolkit and was omitted. Number of words in these six corpuses is about 63 000. Besides that we have two different compiled wordlists. The wordlists are 3850_L (word count 3855), and 3850L_8000M (word count 11 971). 3850_L has been compiled in Department of Modern Languages at the University of Helsinki. 3850L_8000M is a mixture of the data of Crohns and Sundell [12] with 8116 words from the crowd-sourced data combined with the 3850_L wordlist. Both of these lists have word pairs where one is the misspelled word and the other the correct version. The accuracy of the lists has not been intellectually checked, they are used on as is basis.

Some comments on the nature of the evaluation data are in order. Newspaper data is realistic in its error counts, and the six different corpuses have different number of errors, as shown in Table 1. Word pair lists are more artificial in their distributions. 3850_L word list has an error percentage of about 17 % (3195 correct word pairs and 660 erroneous ones), which seems low compared to our real data. 3850L_8000M contains 3393 correct word pairs (72 % errors).

2.3 Evaluation Measures

There are various possible measures to use in OCR post-correction evaluation. In our quality assessment of the Digi data [4] we used word accuracy, word error rate, precision, recall and Fmean for measuring

the number of errors the evaluation samples have. We used four different readymade software for the analysis. Two of them were dedicated OCR evaluation software, two MT quality evaluation software. One of them, OCR Frontiers Toolkit 1.0[1], which measures word accuracy, is also used in this paper because the software is able to evaluate the parallel newspaper data comparing the original data to output of the spelling correction with one word per line. Word level accuracy is not a very good measure while it is only sensitive to the number of errors in comparison and does not show details of correction [14: 269]. With this material, however, it is the most suitable available measure, as the data is not in one-to-one correspondence on word level.

For the wordlist data we have compiled later we use recall, precision and F-score [14: 268–269]. Given that we have tuples of error, original and correction, `<$1, $2, $3>`, we can define true positives, false positives, false negatives and true negatives as follows using Gnu-AWK's notation of *is equal to* `(==)`, *is not equal to* `(!=)` and *conjunction* `(&&)`:

- `(($1 != $2) && ($2 == $3))` **TP**, true positive: a wrongly spelled word is corrected
- `(($1 == $2) && ($2 != $3))` **FP**, false positive: a correct word is changed to a misspelling
- `(($1 != $2) && ($2 != $3))` **FN,** false negative: a wrongly spelled word is wrong after correction
- `(($1 == $2) && ($2 == $3))` **TN**, true negative: a correct word is correct after correction

Recall, R, is TP / (TP+FN), Precision, P, is TP / (TP+FP) and F-score, F, is 2*R*P / (R + P).

2.4 Correction Algorithm

After initial trials with different correction approaches in [4] we have been working with a Levenshtein distance (LD) correction algorithm introduced in [15]. The original version is a Python program that uses only LD 2, so it is able to correct two errors per word at maximum.

---

[1] https://code.google.com/p/isri-ocr-evaluation-tools/

This is a reasonable limit, while many of the OCR errors are in this range. We use the Gnu-AWK (GAWK) version of the algorithm which was implemented by Gregory Greffenstette[2] with some modifications of our own. Levensthein distance, also known as minimum edit distance, is the minimum number of editing operations necessary to transform one word into another. An editing operation is a character insertion, deletion, substitution or transposition.

The original algorithm uses a frequency dictionary as a language model (LM) and makes corrections according to the model. We added another, much larger dictionary to the algorithm to verify first, that the word being processed is not already included in the lexicon and thus possibly a correct spelling. If it is not, the word will be sent to correction. We'll call this dictionary the verification dictionary (VD). We also added one simple rule, change of *c* to *e* (c → e) between non-vowels, as this is one of the most common OCR errors in the data. Some trash deletion was also added, but the core algorithm is the original GAWK implementation. The algorithm returns only the correction, not a list of correction candidates. If the length of the processed word is less or equal to three characters, correction will not be tried in our version. The dictionaries we use with the algorithm have been compiled from different sources using for example frequency list of Early modern Finnish from Kotus[3] with about 530 000 words, four dictionaries from the 19th century[4] and other available material, also from the Digi collection. We have been experimenting with different lexicons and different LD levels with the algorithm, and will report the results in the following.

3 Results

Results of the newspaper data and wordlists are shown and discussed separately as they use different evaluation measures. Table 1 shows results of the newspaper material. We have tried different Levenshtein distance levels from the basic 2 up to 5, but report only the basic results and the results with LD 5, as there is no real difference in most of the cases between the different LD levels.

[2] http://awk.info/?doc/tools/spellcheck.html
[3] http://kaino.kotus.fi/sanat/taajuuslista/vns_frek.zip
[4] http://kaino.kotus.fi/korpus/1800/meta/1800_coll_rdf.xml

Collection	Original word accuracy results from [4]	Correction results with LD 2	Correction results with LD 5	Best correction result vs. original +/-, per cent units
Suometar 1847	71.0 %	79.2 %	79 %	+8.2
Keski-Suomi 1871	60.5 %	70.7 %	70.1 %	+10.2
Sanan Saatta- ja Wiipurista 1841	73.8 %	80 %	79.6 %	+6.2
Turun Wiikko-Sanomat 1831	80.4 %	80.5 %	80.6 %	+0.2
Oulun Viikko-Sanomia 1841	83 %	83.2 %	82.9 %	+0.2
Kirjallinen Kuukauslehti 1870	82.1 %	76.8 %	76.6 %	-5.3

Table 1. Correction results of the newspaper material

Results of the newspaper material correction show that with lower quality data the correction algorithm works reasonably well, it is able to improve word accuracy with 6–10 % units in all three evaluation data sets. With better quality data the results are not that good: correction is able to keep the quality of the data at the same level in two cases, but in one case the quality deteriorates quite a lot, 5.3 % units. Overall the results are fair, but it seems that there is not much possibility for improvement with the used algorithm. Selection of dictionaries used with the algorithm has a modest impact on the results, but it seems that the best results are achieved when the LM dictionary is quite small, about 1.35 M words. Much larger LM dictionaries do not seem to give any gain in performance. Effect of the VD dictionary will be discussed with word list results.

Results of the word list correction are shown in Tables 2 and 3. Table 2 shows results of the 3850_L wordlist, Table 3 shows results of the 3850L_8000M list.

3850_L	**Basic LM LD 2**	**LM W/V LD 2**	**Basic LM LD 5**	**LM W/V LD 5**
With VD	R = 0.43 P = 0.86 F = 0.57 FP = 46	R = 0.44 P = 0.90 F = 0.59 FP = 31	R = 0.44 P = 0.85 F = 0.58 FP = 52	R = 0.49 P = 0.77 F = 0.60 FP = 97
Results without punctuation and numbers	R = 0.42 P = 0.78 F = 0.55 FP = 64	R = 0.44 P = 0.82 F = 0.57 FP = 51	R = 0.44 P = 0.77 F = 0.56 FP = 70	R = 0.49 P = 0.69 F = 0.58 FP = 116
Without VD	R = 0.47 P = 0.74 F = 0.58 FP = 109	R = 0.49 P = 0.77 F = 0.60 FP = 97	Same as in column 2	Same as in column 3
Results without punctuation and numbers	R = 0.48 P = 0.66 F = 0.56 FP = 127	R = 0.49 P = 0.69 F = 0.58 FP = 116	Same as in column 2	Same as in column 3

Table 2. Correction results of the 3850_L word list

Table legend: R = recall, P = precision, F = F-score, FP = number of false positives, LD2 and LD 5 = Levenshtein edit distances of 2 and 5, LM W/V = language model dictionary contains both v and w versions of words that have either. v/w variation is one of the basic features of 19th century Finnish orthography.

There are some clear and mainly expected trends in the results. Usage of the VD, verification dictionary, improves precision and hurts recall to some extent. This can also be seen in the number of false positives, which doubles or almost triples if no lexical check is done before sending the word to correction. Size of the VD is 4.96 M words.

Recall in the 3850_L sample varies from about 0.44 to 0.49. Precision varies from 0.66 to 0.90, and F-score is round 0.55–0.59. In the 3850_8000M sample recall varies from 0.27–0.34 and precision from 0.89–0.97, F-score being from 0.42 to 0.50. Language model dictionary

that has both *v* and *w* versions of words containing either letter improves recall with about 1.0 % unit and precision with 2–3 % units. Punctuation and numbers do not occur much in the 3850_L sample and their inclusion or exclusion in the evaluation does not change results. In the 3850_8000M sample results without punctuation and numbers are about 6–8 % units better than with punctuation and numbers.

We can see that usage of LD 5 does not improve results much. Recall can be slightly better when using LD 5, but precision is worse with 3850_L and at the same level with 3850_8000M. Usage of the VD is not clearly beneficial with the wordlists, although it gave the best results with the newspaper material. This may be due to different measures: word accuracy hides details of performance, and the improvement VD brings with the newspaper material is shown to be more ambiguous when precision and recall are used as measures.

3850L_8000M	**Basic LM, LD 2**	**LM W/V, LD 2**	**Basic LM, LD 5**	**LM W/V, LD 5**
With VD	R = 0.28 P = 0.95 F = 0.43 FP = 118	R = 0.28 P = 0.96 F = 0.43 FP = 108	R = 0.27 P = 0.97 F = 0.42 FP = 77	R = 0.27 P = 0.97 F = 0.43 FP = 67
Results without punctuation and numbers	R = 0.34 P = 0.92 F = 0.50 FP = 243	R = 0.35 P = 0.92 F = 0.50 FP = 239	R = 0.34 P = 0.93 F = 0.49 FP = 205	R = 0.34 P = 0.93 F = 0.50 FP = 201
Without VD	R = 0.28 P = 0.93 F = 0.42 FP = 173	Same as in column 2	Same as in column 2	Same as in column 2
Results without punctuation and numbers	R = 0.34 P = 0.89 F = 0.49 FP = 331	Same as in column 2	Same as in column 2	Same as in column 2

Table 3. Correction results of the 3850L_8000M word list

4 Discussion and Conclusion

We have reported in this paper first results of post-correction of OCRed data from a Finnish digitized newspaper and magazine collection, that

contains 1.95 M pages of data and about 837 million words of Finnish from the period between 1771 and 1910. Our sample evaluation data are mainly from years between 1830 and 1870 and year 1882, which is the period of so called Early modern Finnish. Evaluations of the post-correction were partly based on parallel text material gathered earlier [4] and partly on compiled word pair lists of the digitized material. The chosen post-correction method was a straightforward Levenshtein distance based algorithm with some additions.

The results we have shown are fair, but not good enough for realistic post-correction as the only correction method. However, they show that the quality of even quite poor OCR output can be improved with a quite simple approach. If the data were not so bad, we could perhaps be able to improve the quality of the Digi collection even with the current quite simple approach enough for our purposes. Our material contains lots of hard errors, and as it was seen in the results section, only the simplest ones seem to get corrected and usage of deeper LD levels does not help. Usage of the VD dictionary helps in correction, but increasing its size substantially does not bring much improvement. Without VD look-up the correction algorithm creates quite a lot of false positives that decrease precision. Size of the LM dictionary (1.35 M tokens) seems quite optimal. Including the v/w variation in the LM dictionary seems to be beneficial, too.

Many of the OCR post-correction evaluations use data that has already a quite high correctness percentage [6] and thus they can also set high expectations for the results achieved. Examples of our data, the British Library data [10] and The Times of London [8] show that the quality level of a large OCRed 19th century newspaper collection is not very high and thus it is only reasonable to set the aims in correction not too high. If we can improve the quality of the data with usage of re-OCRing and isolated word post correction cycle to the level of some 80 – 90 % word correctness overall, that would improve usability of the material a lot, and would also meet our current needs quite well. After that context sensitive real world spelling correction might also be possible, if that would be needed.

The main value of our work so far has been the set-up of the whole correction and evaluation chain and gaining experience with the correction and the data. We have acquired invaluable experience concerning the quality of our material and gathered both useful tools and word list

data to be used in the correction. With the experience we can plan further steps of the post-correction realistically.

Acknowledgments

This research is funded by the EU Commission through its European Regional Development Fund, and the program *Leverage from the EU 2007–2013.*

References

1. Bremer-Laamanen, M.-L.: A Nordic Digital Newspaper Library. International Preservation News 26, 18–20 (2001)
2. Bremer-Laamanen, M.-L.: Connecting to the Past – Newspaper Digitization in the Nordic Countries. World Library and Information Congress. In: 71th IFLA General Conference and Council, "Libraries - A voyage of discovery", August 14th - 18th 2005, Oslo, Norway (2005) Available at http://archive.ifla.org/IV/ifla71/papers/019e-Bremer-Laamanen.pdf
3. Bremer-Laamanen, M.-L.: In the Spotlight for Crowdsourcing. Scandinavian Librarian Quarterly 1,18–21 (2014)
4. Kettunen, K., Honkela, T., Lindén, K., Kauppinen, P., Pääkkönen, T., Kervinen, J.: Analyzing and Improving the Quality of a Historical News Collection using Language Technology and Statistical Machine Learning Methods. In: Proceeding of IFLA 2014, Lyon (2014) http://www.ifla.org/files/assets/newspapers/Geneva_2014/s6-honkela-en.pdf
5. Holley, R.: How good can it get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. D-Lib Magazine 3 (2009) http://www.dlib.org/dlib/march09/holley/03holley.html
6. Furrer, L., Volk, M.: Reducing OCR Errors in Gothic-Script Documents. In: Proceedings of Language Technologies for Digital Humanities and Cultural Heritage Workshop, 97–103, Hissar, Bulgaria (2011)
7. Klijn, E.: The Current State-of-art in Newspaper Digitization. A Market Perspective. D-Lib Magazine 14 (2008.) http://www.dlib.org/dlib/january08/klijn/01klijn.html
8. Niklas, K.:. Unsupervised Post-Correction of OCR Errors. Diploma Thesis, Leibniz Universität, Hannover (2010) www.l3s.de/~tahmasebi/Diplomarbeit_Niklas.pdf
9. Taghva, K., Borsack, J., Condit, A.: Evaluation of Model-Based Retrieval Effectiveness with OCR Text. ACM Transactions on Information Systems, 14, 64–93 (1996)
10. Tanner, S., Muñoz, T., Ros, P. H.: Measuring Mass Text Digitization Quality and Usefulness. Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive. D-Lib Magazine 15 (2009) http://www.dlib.org/dlib/july09/munoz/07munoz.html
11. Lopresti, D.: Optical Character Recognition Errors and their Effects on Natural Language Processing. International Journal on Document Analysis and Recognition 12, 141–151 (2009)

12. Chrons, O., Sundell, S.: Digitalkoot: Making Old Archives Accessible Using Crowdsourcing. In: Human Computation, Papers from the 2011 AAAI Workshop (2011) http://www.aaai.org/ocs/index.php/WS/AAAIW11/paper/view/3813/4246.
13. Kukich, K.: Techniques for Automatically Correcting Words in Text. ACM Computing Surveys 24, 377–439 (1992)
14. Manning, C. D. , Schütze, H.: Foundations of Statistical Language Processing. The MIT Press, Cambridge, Massachusetts (1999)
15. Norvig, P.: How to Write a Spelling Corrector (2008) norvig.com/spell-correct.htm