Research and development efforts on the digitized historical newspaper and journal collection of The National Library of Finland

Kimmo Kettunen [0000-0003-2747-1382] , Mika Koistinen and Teemu Ruokolainen

The National Library of Finland, DH projects Saimaankatu 6, 50 100 Mikkeli, Finland
firstname.lastname@helsinki.fi

Abstract. The National Library of Finland (NLF) has digitized historical newspapers, journals and ephemera published in Finland since the late 1990s. The present collection consists of about 12.8 million pages mainly in Finnish and Swedish. Out of these about 7.36 million pages are freely available on the web site digi.kansalliskirjasto.fi (Digi). The copyright restricted part of the collection can be used at six legal deposit libraries in different parts of Finland. The time period of the open collection is from 1771 to 1929.

This paper presents work that has been carried out in the NLF related to the historical newspaper and journal collection. We offer an overall account of research and development related to the data.

Keywords: historical newspaper collections, research and development, Finnish

1 Introduction

Digitization of historical newspapers and journals printed in Finland began at the end of the 1990s in The National Library of Finland. Besides producing the digitized publication data all the time NLF has been involved in research and improvement of the digitized material during the last years. Last summer we ended a two year European Regional Development Fund (ERDF) project and started another two year ERDF project in August 2017. NLF is also involved in research consortium COMHIS, *Computational History and the Transformation of Public Discourse in Finland, 1640-1910*, that is funded by the Academy of Finland (2016–2019). COMHIS utilizes NLF's historical newspaper and journal data as one of its main sources in its research of changes of publicity in Finland.

So far we have focused on text material, and our main achievements have been the following: a thorough quality analysis of the Finnish part of the textual data on word level [1], evaluation of tools for named entity recognition and setting up of a NER evaluation collection [2] and an improved Optical Character Recognition framework for the data using Tesseract open source OCR engine [3]. In March 2017 we published the newspaper and journal text data as an open data collection on our web pag-

es [4]. We have also started work on article extraction from the pages of one newspaper – Uusi Suometar: 1869 –1918, 86 068 pages.

Newspapers and journals contain also lots of pictures: drawings, maps, photographs and different graphs which are of interest to both researchers and lay persons. So far we can detect pictures on the pages of digi.kansalliskirjasto.fi, but a systematic way of both detecting the pictures and classifying their content needs to be developed. This is one of our aims during the two year ERDF funding period.

2 Research and development

We shall describe our efforts with the newspaper and journal texts in general for the rest of the paper. Those who are interested in more details can read the published research papers for each topic.

2.1 Quality analysis and a new OCR process

When originally non-digital materials, such as old newspapers and books, are digitized, the process starts with scanning of the documents which results in image files. Out of the image files one needs to sort out texts and possible non-textual data, such as photographs and other pictorial representations. Texts are recognized from the scanned pages with Optical Character Recognition (OCR) software. OCRing for modern text types and fonts is considered as a resolved problem, that yields high quality results, but results of historical document OCRing are still far from that. Scanned and OCRed document collections have usually a varying amount of errors in their content. Tanner et al. [5], for example, report that 78% of the words in the collection of *The 19th Century Newspaper Project* of the British Library are correct. This quality is not good, but quite common to many comparable collections [6].

To be able to improve the data of our collections we needed first to get an overall impression of its quality. We have performed a comprehensive general analysis for the Finnish data of 1771–1910 [1]. Out of the assessment we have learned that about 70–75% of the words in the Finnish data are probably right and recognizable. In a collection of about 2.4 billion words this means that about 600–800 million word tokens and about 200 million word types are wrong.

In order to improve the quality of the collection, we started to consider re-OCRing of the data in 2015. The main reason for this was that the collection had been OCRed with a proprietary OCR engine, ABBYY FineReader (v.7 and v.8). Newer versions of the software exist, the latest being 14.0, but the cost of the Fraktur font for OCR is too high a burden for re-OCRing the collection with ABBYY FineReader. We ended up using open source OCR engine Tesseract v. 3.04.01[1] and started to train Fraktur font for it. This process and its results are described in detail in Koistinen et al. [3]. We have an evaluation collection for the re-OCR, and evaluation results show that Tes-

[1] https://github.com/tesseract-ocr

seract OCR words are recognized by morphological analyzer 9% units better than words of present OCR (90% vs. 81 % in the evaluation collection).

We have performed a small scale re-OCR with our real data. With 7 937 pages of Uusi Suometar 1869–1879 morphological analyzer recognized 73% of ca. 13.45 M words of old OCR. After re-OCR 86.5% of the words were recognized, which is a 13.5% unit improvement in recognition. The data used represents a realistic printed newspaper of its time with a varying column layout.

2.2 Named entity recognition

Digital collections of the NLF are part of the growing global network of digitized newspapers and journals, and historical newspapers are considered more and more as an important source of historical knowledge. As the amount of digitized journalistic information grows, also tools for harvesting the information are needed. Named Entity Recognition (NER) has become one of the basic techniques for information extraction of texts since the mid-1990s [7]. In its initial form NER was used to find and mark semantic entities like person, location and organization in texts to enable information extraction related to these kinds of entities. Later on other types of extractable entities, like time, artefact, event and measure/numerical, have been added to the repertoires of NER software [7].

We started by performing preliminary NER analysis of our data with a variety of tools in an evaluation collection [2]. Results of the evaluation showed that bad OCR quality harms NER clearly, and also the tools that we had at use were not the best possible, as they were analyzers of modern Finnish. Anyhow, we were able to achieve F-score of about 0.5–0.6 at best with persons and locations. Organizations were not recognized as well.

After the initial evaluation we have set up a larger evaluation and training collection. We have been able to train a standard statistical NER package, Stanford NER, with our data. Our initial results with the Stanford NER were mainly promising: persons achieved F-score of 0.69, locations F-score of 0.74, and organizations F-score of 0.36 in the re-OCRed evaluation collection. With more training data – final word account is 381 356 words – the results have improved: locations achieve now F-score of 0.79, persons 0.72 and organizations 0.42. Results for locations and persons can be considered quite good and the resulting NER model will be used in our web collection to improve access to the content.

2.3 Article extraction

It is common that historical newspaper collections are digitized on page level: pages of the physical newspapers are scanned and the page images serve also as the basic browsing and searching unit of the collection. Searches to the collection are performed as well as results are shown to the user on page level. Page, however, is not the basic informational unit of a newspaper, but pages consist of articles, although length of articles can be quite variable. Thus separation of the article structure of the digitized newspaper pages is an important step to improve usability of the digital col-

lections. As Dengel and Shafait [8] formulate it: “availability of logical structure facilitates navigation and advanced search inside the document as well as enables better presentation of the document in a possibly restructured format.” Article structure will also enhance further analysis stages of the content, such as topic modelling or any other kind of content analysis. Information retrieval performance of the newspaper collection should also improve, if its content were indexed on article level [9]. Several digitized historical newspaper collections have implemented article extraction on their pages. Good examples are for example Italian La Stampa, The British Newspaper Archive, and Australian Trove.

Article extraction on digitized newspaper pages is not an easy task. Results of the biannual ICDAR competition on historical newspaper layout analysis show that current algorithms segment and label about 80–85% of the pages correctly at best [10].

We have so far investigated different possibilities for article extraction. Available commercial solutions are either too expensive, not good enough or need too much post correction for the articles. We have been evaluating one research software, PIVAJ, that has been developed at the LITIS laboratory of University of Rouen [11]. We intend to use the software for article extraction

Acknowledgements

This work is funded by the European Regional Development Fund and the program Leverage from the EU 2014–2020.

References

1. Kettunen, K., Pääkkönen, T.: Measuring Lexical Quality of a Historical Finnish Newspaper Collection – Analysis of Garbled OCR Data with Basic Language Technology Tools and Means. In: Calzolari, N et al. (eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016) http://www.lrec-conf.org/proceedings/lrec2016/pdf/17_Paper.pdf (2016).
2. Kettunen, K., Mäkelä, E., Ruokolainen, T., Kuokkala, J., Löfberg, L.: Old Content and Modern Tools – Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771–1910. Digital Humanities Quarterly. http://www.digitalhumanities.org/dhq/vol/11/3/000333/000333.html (2017).
3. Koistinen, M., Kettunen, K., Pääkkönen, T.: Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur & Antiqua Models and Image Preprocessing. In: Proceedings of Nodalida 2017 http://www.ep.liu.se/ecp/131/038/ecp17131038.pdf (2017).
4. Pääkkönen, T., Kervinen, J., Nivala, A., Kettunen, K., Mäkelä, E.: Exporting Finnish Digitized Historical Newspaper Contents for Offline Use. D-Lib Magazine, July/August. http://www.dlib.org/dlib/july16/paakkonen/07paakkonen.html (2016).
5. 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, (15/8) http://www.dlib.org/dlib/july09/munoz/07munoz.html (2009).

6. Jarlbrink, J., Snickars, P.: Cultural heritage as digital noise: nineteenth century newspapers in the digital archive. *Journal of Documentation*, https://doi.org/10.1108/JD-09-2016-0106 (2017).
7. Nadeau, D., Sekine, S.: A Survey of Named Entity Recognition and Classification. Linguisticae Investigationes 30, 3–26 (2007).
8. Dengel, A., Shafait, F.: Analysis of the Logical Layout of Documents. In Doerman, D., Tombre, K. (eds.) Handbook of Document Image Processing and Recognition, 177–222. Springer. DOI 10.1007/978-0-85729-859-1 (2014).
9. Buhr, F., Neumann, B.: Evaluation of Retrieval Performance in Historical Newspaper Archives comparing Page-level and Article-level Granularity. https://kogs-www.informatik.uni-hamburg.de/publikationen/pub-buhr/Newspaper-Retrieval.pdf (2014).
10. Antonacopoulos, A., Clausner, C., Papadopoulos C., Pletshacher, S.: ICDAR2013 Competition on Historical Newspaper Layout Analysis – HNLA2013. DOI: 10.1109/ICDAR.2013.293 (2013).
11. Hebert, D., Palfray, T., Nicolas, T., Tranouez, P., Paquet, T.: Automatic article extraction in old Newspapers Digitized Collections. In Proceeding DATeCH '14 Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, 3–8. http://dl.acm.org/citation.cfm?id=2595195 (2014).