Reusing the Model and Components of an IIR Study for Perceived Effects of OCR Quality Change

Kimmo Kettunen (UEF)
Heikki Keskustalo (Tuni)
Birger Larsen (AAU)
Tuula Pääkkönen (NLF)
Juha Rautiainen (NLF)

Optical Character Recognition quality and Information Retrieval from Historical Newspapers

- Digitized historical newspaper collections have been produced and increasingly used during the last two decades in different parts of the world, and their usage and demand will increase in the future. Access to these collections is important for different user groups, such as lay persons, teachers, journalists, and professional historians.
- Contents of the historical newspaper collections are produced using Optical Character Recognition, which has produced results of varying quality in the past.
- It is known that OCR noise present in digitized historical documents disturbs end user perception of documents. However, this effect on the desired access is difficult to study.
- In this paper, we describe a research design intended to allow studying this issue and **discuss the reusability of the model and its components**

Optical Character Recognition quality and Information Retrieval

Effects of low OCR quality have been shown in laboratory style IR studies for a long time, cf. e.g.

- **Paul B. Kantor, Ellen M. Voorhees, E.M. 2000**. The TREC-5 confusion track: comparing retrieval methods for scanned text. Inf. Retrieval 2(2), 165–176; **Guilherme Torresan Bazzo, Gustavo Acauan Lorentz, Danny Suarez Vargas, and Viviane P. Moreira. 2020.** Assessing the Impact of OCR Errors in Information Retrieval. In Jose J. et al. (eds) Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science, vol 12036. Springer, Cham. DOI: https://doi.org/10.1007/978-3-030-45442-5_13; **Anni Järvelin, Heikki Keskustalo, Eero Sormunen, Miamaria Saastamoinen, and Kimmo Kettunen. 2016**. Information retrieval from historical newspaper collections in highly inflectional languages: A query expansion approach. J. Assoc. Inf. Sci. Technol. 67, 12 (December 2016), 2928–2946. DOI: https://doi.org/10.1002/asi.23379.

Also, digital humanities scholars have been complaining, cf. e.g.

- **Johan Jarlbrink and Pelle Snickars. 2017**. Cultural heritage as digital noise: nineteenth century newspapers in the digital archive. Journal of Documentation 73: 1228-1243. DOI: 10.1108/JD-09-2016-0106; Eva **Pfanzelter, Eva, Sarah Oberbichler, Jani Marjanen, Pierre-Carl Langlais, and Stefan Hechl. 2021.** Digital interfaces of historical newspapers: opportunities, restrictions and recommendations. Journal of Data Mining and Digital Humanities. DOI: 10.46298/jdmdh.6121.

No controlled user studies so far!

- Besides our current study there does not seem to exist any real user study related to the effects of OCR quality with historical newspapers
- Our study is a user study using a task-based interactive information retrieval (TBII) model
- Paper related to the study published in IRDCL2022: ”OCR quality affects perceived usefulness of historical newspaper clippings – a user study” http://ircdl2022.dei.unipd.it/downloads/papers/IRCDL_2022_paper_2.pdf

The setting: optically read historical newspaper content with automatically induced article structure

- Contents of one Finnish newspaper, Uusi Suometar 1869-1918: ca. 86 000 pages, and 306 million words
- Content separated automatically to ”articles” with PIVAJ (from LITIS lab, Rouen)
- Available two different OCR qualities for texts: original and improved - difference in word recognition ca. 15%-units (measured with automatic morphological analysis)
- Same article structure in both OCR qualities, retrieval all the the time in the better quality index
- Question: do users perceive the quality difference, when they accomplish task-based information retrieval (without knowing about the quality differences)

The TBII setting

- 30 topics with short pre-defined queries were created for the timeline of 1870-1918
- 32 users were recruited to make searches to the Uusi Suometar database
- Each user performed six (6) queries and evaluated the top-10 results for each query
- Task-based background story

”Imagine that you are writing an article related to topics in history of Finland or world history at the end of 19th century or the beginning of 20th century. Evaluate quality of the clippings you get as search results. Evaluate the quality of each clipping from the viewpoint, how it helps you to proceed with your article writing.”

- Evaluation scale of 0-3 used

30 topics with ready made short queries.

The query interface is on the right

Results: a statistically significant difference found

Mean averages for the evaluation scores over the whole query set for pre-formulated queries for the old OCR was 1.26 and for the new OCR 1.36. This reveals that the query results benefited from the improved optical character recognition. The mean average evaluation score for the improved OCR query results is 7.94% higher than the mean average score of the old OCR query results.

The difference in the effect of Optical Character Recognition quality on the relevance judgements was statistically significant ($p=0.002$, Wilcoxon's signed rank test [24]), when the relevance of the individual underlying documents was judged based on two possible levels of Optical Character Recognition quality. The difference in the overall effectiveness of retrieval (measured with mean average of cumulated gain among top-10 documents in the case of 30 topics), however, was not statistically significant ($p=0.10$, Wilcoxon's signed rank test).

Resources and their re-use (Gäde et al. 2021)

- In the following we use the resource type classification presented in Gäde et al. 2021 and divide the resources of the IIR retrieval system to three types:
- 1) research design
- 2) research infrastructure, and
- 3) research data.

In Gäde et al. *research design* is defined "as methods and techniques used to collect and analyse empirical data."

- *Research infrastructure* is defined mainly in relation to the technical infrastructure that is needed to carry out an interactive information retrieval study.
- *Research data* can be broadly defined as "any data that has been collected, observed, generated or created during or as results of the research process" .
- Gäde et al. discuss the notion of *reusability* with regards to different current research articles. After some discussion they define a broad sense of reusability for the IIR community: reuse is "use of research data, research design or infrastructure *for more than an individual purpose*

Parts of our IIR research along Gäde et al's views

Research design

- **Topics and queries** (30 topics)
- **Simulated task** ("write an article")
- **Recruitment of participants** (students, 32)
- **The data collection protocol** (evaluation measures 0-3)
- **Analysis methods** (evaluation scores for relevance of top-10 hits 0-3)
- **Evaluation measures and significance tests** (mean averages for evaluation scores)

Research infrastructure

- **Target data** (contents of one newspaper 1869-1918)
- **OCR software and OCR quality** (two qualities with clear difference)
- **Segmentation software** (PIVAJ machine learning → articles))
- **Search software and search index** (Elastic search, lemmatized index)
- **The query interface** (outsourced, simple)
- **User management and interaction logging** (an Excel sheet with 11 elements)

Research data

- **Research data -** user sessions

How reusable are our research components I?

- **Research design**
  - **The topic descriptions** can be shared via public repositories; the pre-formulated queries are reported in Kettunen et al. [19]. This component can be adapted by creating new topics and queries as necessary, acknowledging the type of target data and its expected use [5, 24]. In the historical newspaper context, it might also be advantageous if a professional historian would take part in topic creation.
  - **Simulated task:** a well-known res. design method; the component can be adapted by creating variations of the specific tasks described by Kettunen et al. [19] in corresponding settings, possibly with the help of professional historians.
  - **Recruitment of participants**: we had mainly students of IR science, but perhaps professional historians would be better → can you get a large enough group?
  - **The Data Collection Protocol:** very simple in our study, could be augmented
  - **Analysis Methods:** we used only means of evaluation scores → could be augmented with other measures
  - **Evaluation Measures and Significance Tests:** common measures and tests, can be adapted/augmented

In General

- All these can be re-used or adapted without too many problems

How reusable are our research components II?

Research Infrastructure

Target data: the Finnish newspaper collection is out of © and thus reusable BUT: the collection is out of general interest due to the language

Software components: OCR software, article extraction software, search engine → In general, maintaining and reusing realistic software components is a hard problem [25]. In our case, e.g., we have utilized NLF's document presentation system, different OCR software, an experimental page segmentation software and a search engine. The query interface and user logging needed to be adapted into this environment. Adapting these components for reuse requires cooperation with the software maintainers or providers in practice.

The Query Interface: one of the most important elements of an IIR study: ours is quite simple and made specifically for this use → could be used as a model

User Management and Interaction Logging: minimal amount of information gathered, no GDPR problems → could be improved

How reusable are our research components III?

Research Data - User Sessions

- The results of the experiment are research data that was accumulated in the search sessions. From the search sessions of one user group 3983 evaluations for query results were collected using the specified log format. From the point of view of the reuse, the gathered data from user sessions is of high value, as it enables the analysis of the query sessions from many viewpoints. So far only basic analysis of the results with pre-formulated queries have been reported in Kettunen et al. [19].
- The result data does not include any personal information, that would hinder its publication. Thus, it is reusable.

So, where are we with reuse?

- The **general model** developed in the study is reusable, but the specific components of it cannot be reused easily.
- Even if there are not many proprietary parts in the overall system, the intertwining and combining of several components makes system's reuse outside of The National Library of Finland hard.
- Some of the components - **target data, topics and pre-formulated queries, and the resulting research data** - however, could be made publicly available and reused, if needed.
- Some of the components - simulated tasks, analysis, and evaluation methods - on the other hand, are general working methods used in IIR. They can be either quite easily reused as such or remodified for new studies.

Conclusion

- **To sum up**, the research design presented is reusable as a whole or as a **model.**
- However, the adaptation of each individual component of the research setting must be considered when the reuse scenario is planned.
- Some components of the design, such as recruiting of participants, may be relatively straightforward to replicate in a new study; adapting other components, such as planning user tasks which allow focusing on, e.g., selected user activities or modifying software, may be more laborious and often require specific expertise.

NEWS EYE

A Digital Investigator for Historical Newspapers

AALBORG UNIVERSITY
DENMARK

UNIVERSITY OF EASTERN FINLAND

Thank you!

Tampereen yliopisto
Tampere University

1640
NATIONAL LIBRARY OF FINLAND

Kimmo Kettunen (UEF)

Heikki Keskustalo (Tuni)

Birger Larsen (AAU)

Tuula Pääkkönen (NLF)

Juha Rautiainen (NLF)