Improving OCR of historical newspapers and journals published in Finland by adding Swedish training data

Senka Drobac
Pekka Kauppinen
Krister Lindén

What did we do?

In previous work (Drobac 2017) we trained Finnish models for Optical Character Recognition (OCR) of historical newspapers and journals published in Finland. While those models are doing reasonably well on Finnish test set with 95.21% CAR, they are well below 90% CAR on Swedish test sets. Since large amount of historical newspapers and journals published in Finland was written in Swedish, we need to improve our models in order to perform good quality OCR on the entire corpus.

Here we show test results of experiments performed on Finnish and Swedish models as well as some mixed models. All models are trained with Ocropy1 toolkit. Tests show that we get the best results on Finnish test sets by adding a small amount of Swedish data to the Finnish data. On Swedish test sets, we still get poor results, but that could be due to the small amount of the Swedish data that we currently have. Since this is still work in progress, we are working on acquiring more training data, both for Swedish and Finnish written in Antiqua typeface.

[1]**Ocropy** - leading open source software toolkit for OCR, uses long short term memory networks and comes with document pre-processing tools

Motivation

OCR of historical texts is difficult because of:

- font diversity
- lack of orthographic standard (same words spelled differently)
- material quality (some documents can have deformations)
- a lexicon of known historical spelling variants is not available

Specific problems for this corpora:

- Two typefaces: Fraktur and Antiqua
- Two languages: Finnish and Swedish

ABBYY FineReader's* CAR on our test sets: 90% - 91%
(*commercial software that National Library of Finland used to OCR the corpus)

Evaluation

Evaluation metric that we used for OCR models and correction models is character accuracy rate (CAR) and word accuracy rate (WAR). It is calculated from the Levenshtein distance between system output and ground truth (for full lines):

$$CAR, WAR = 100\% \times \frac{correct}{correct + errors}$$

Results

Test set	FIN MODEL	SWE MODEL
Fin-Fraktur	**95.43 / 78.79**	93.2 / 69.61
Fin-Antiqua	85.81 / 53.36	**88.89 / 62.32**
Swe-Fraktur	78.84 / 40.43	**87.59 / 55.32**
Swe-Antiqua	79.93 / 40.01	**90.66 / 66.36**

Table 1: CAR/WAR (%) for individual models trained on Finnish and Swedish data separately and tested on four test sets: Finnish Fraktur, Finnish Antiqua, Swedish Fraktur and Swedish Antiqua. Finnish model was trained on ~10 000 lines, Swedish model on ~3 300 lines.

Not enough Finnish Antiqua in training

Finnish results are better with some Swedish data in the model

Results on Swedish test sets are still poor, possibly because of small training set

Test set	MODEL FIN + SWE 1	MODEL FIN + SWE 2	MODEL FIN + SWE 3
Fin-Fraktur	**96.19 / 81.91**	95.07 / 76.65	94.97 / 76.13
Fin-Antiqua	**89.35 / 63.35**	87.23 / 58.22	86.64 / 55.79
Swe-Fraktur	**82.53 / 51.11**	80.76 / 43.48	**83.22 / 45.65**
Swe-Antiqua	**86.65 / 59.84**	83.69 / 49.49	84.88 / 52.5

Table 2: CAR/WAR (%) for models trained on (1) ~10 000 Finnish + 840 Swedish lines, (2) ~10 000 Finnish + 1680 Swedish lines, (3) ~10 000 Finnish + 3360 Swedish lines and tested on four test sets: Finnish Fraktur, Finnish Antiqua, Swedish Fraktur and Swedish Antiqua.

Data sets

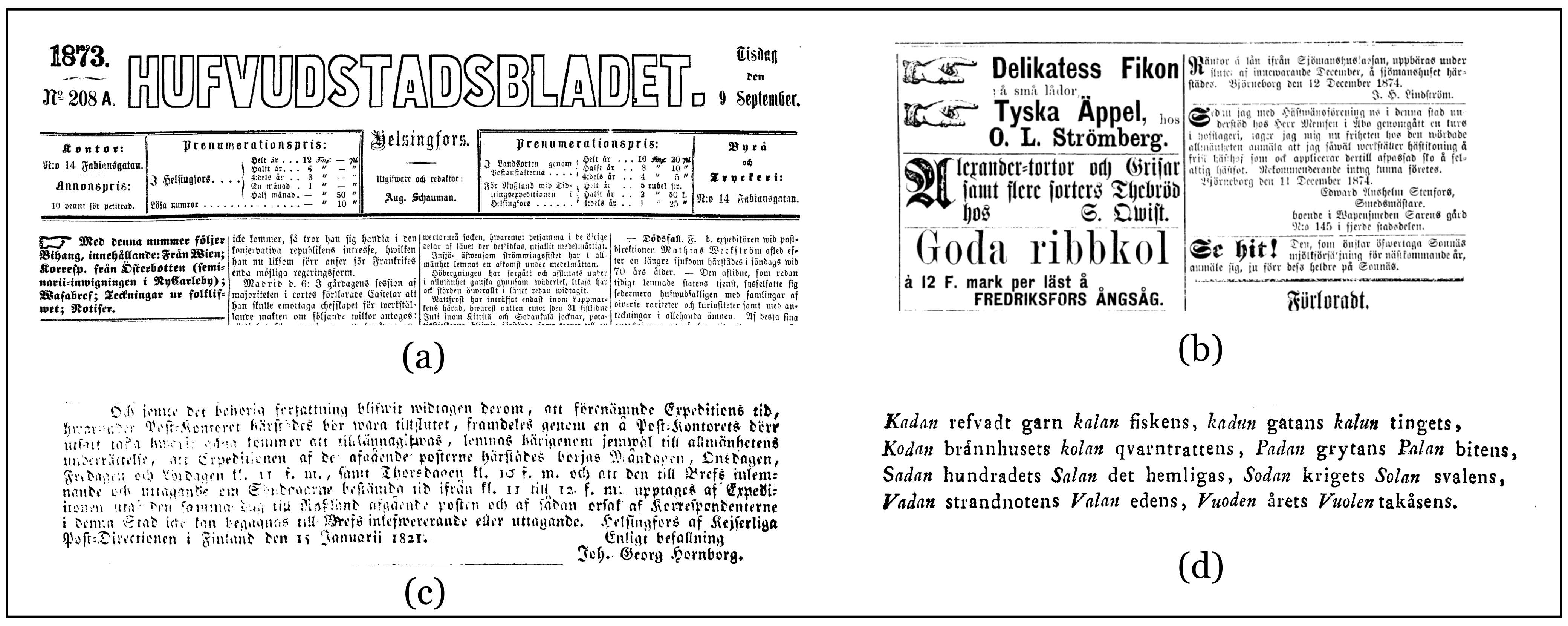
Experiments were done on two data sets created from the historical newspaper and journal corpus compiled by the National Library.

FIN

- ~12 000 random image lines of Finnish text and manually created ground truth
- ~10 000 lines are used for training, the rest is for validation and testing
- Ratio of typefaces: ~75% Fraktur, ~25% Antiqua

SWE

- ~4 000 random image lines of Swedish text and manually created ground truth
- ~3 300 lines are used for training, the rest is for validation and testing
- Ratio of typefaces: ~50% Fraktur, ~50% Antiqua


1873.
№ 208 A.
HUFVUDSTADSBLADET.
Tisdag den 9 September.
Helsingfors.
Prenumerationspris:
Delikatess Fikon
Tyska Äppel,
O. L. Strömberg.
Goda ribbkol
à 12 F. mark per läst å
FREDRIKSFORS ÅNGSÅG.
Förloradt.

(a)

(b)

Och jemte det behöriga författning blifwit widtagen derom, att förenämnde Expeditions tid, hwarunder Poſt-Kontoret härſtädes bör wara tillſlutet, framdeles genom en å Poſt-Kontorets dörr uſatt tafla hwarje gång [illegible] kommer att tillkännagifwas, lemnas härigenom jemwäl till allmänhetens underrättelſe, att Expeditionen af de afgående poſterne härſtädes begynnas Måndagen, Onsdagen, Fredagen och Lördagen kl. 11 f. m., ſamt Thorsdagen kl. 10 f. m. och att den till Brefs inlemnande och uttagande om Söndagarne beſtämda tid ifrån kl. 11 till 12 f. m. upptages af Expeditionen utaf den ſamma dag till Ryßland afgående poſten och af ſådan orſak af Korreſpondenterne i denna Stad icke kan begagnas till Brefs inleſwererande eller uttagande. Helſingfors af Kejſerliga Poſt-Directionen i Finland den 15 Januarii 1821.
Enligt befallning
Joh. Georg Hornborg.

(c)

Kadan refvadt garn *kalan* fiskens, *kadun* gatans *kalun* tingets,
Kodan brännhusets *kolan* qvarntrattens, *Padan* grytans *Palan* bitens,
Sadan hundradets *Salan* det hemligas, *Sodan* krigets *Solan* svalens,
Vadan strandnotens *Valan* edens, *Vuoden* årets *Vuolen* takåsens.

(d)

Image 1: Extracts from 4 binarized pages in the corpus. Images (a) and (b) are examples of difficult segmentation. Image (b) also contains multiple different fonts of both typefaces. Image (c) is an example of poor quality image and image (d) shows use if two languages in one page.

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI
HUMANISTINEN TIEDEKUNTA
HUMANISTISKA FAKULTETEN
FACULTY OF ARTS