Re-OCRing the 19th Century

A New Ground Truth for Training OCR Engines on Novels in Norwegian Fraktur

Andre Kåsen
Language Technologist
Language Bank
National Library of Norway
E-mail: andre.kasen@nb.no

Lars Johnsen
Research librarian
Language Bank
National Library of Norway
E-mail: lars.johnsen@nb.no

National Library of Norway

The Fraktur Problem

In 2006, the National Library of Norway started a digitizing program with the aim of digitizing its entire collection. Material that includes text, such as books, is sent for Optical Character Recognition (OCR), a process that involves computers "translating" pictures into text. However, parts of the older collection has systematic OCR errors. The error addressed in the present project is also known as **the Fraktur problem**. As more robust and scalable technical solutions become available, an effort to correct these errors is called for.

The bar chart below shows the frequency count of OCR errors and spelling variants for the word Læseren (the reader). Here we see that the letter æ is often OCRed as either ce or cr.

The Solution

In this pilot project we seek to explore the new generation of neural systems and establish a new ground truth for OCR-reading Norwegian Fraktur, beginning with works from the mid-to-late 1800's and then moving backward. Currently our training, development and test data consist of the books *En Glad Gut* (1868) by Bjørnstjerne Bjørnson, *Symra* (1863) by Ivar Aasen and *Ferdaminni fraa Sumaren 1860* (1871) by Aasmund Olavsson Vinje.

As the size of the data set grows, we will iteratively use the new ground truth to train models that can be used to process older books that will be manually corrected and then added to the data set. This process will yield both a data set and a set of models that we will make publicly available.


A B C = Bog.

Selvlydende Bogstaver:

i y u a aa æ e o ø.

Medlydende Bogstaver:

j l k t f s r n m

h b d b p g q

ß ſ.

Store Bogstaver:

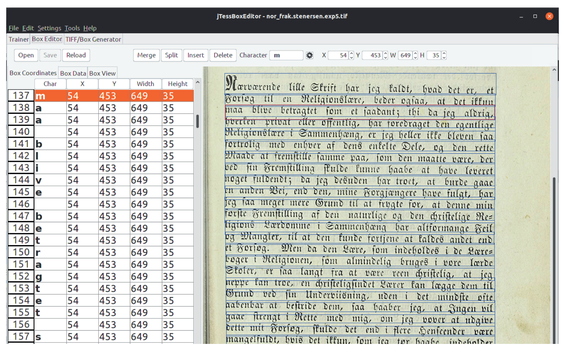
A Æ E C G H K

B V R N M P Y

D Ø Q Ø J F T L

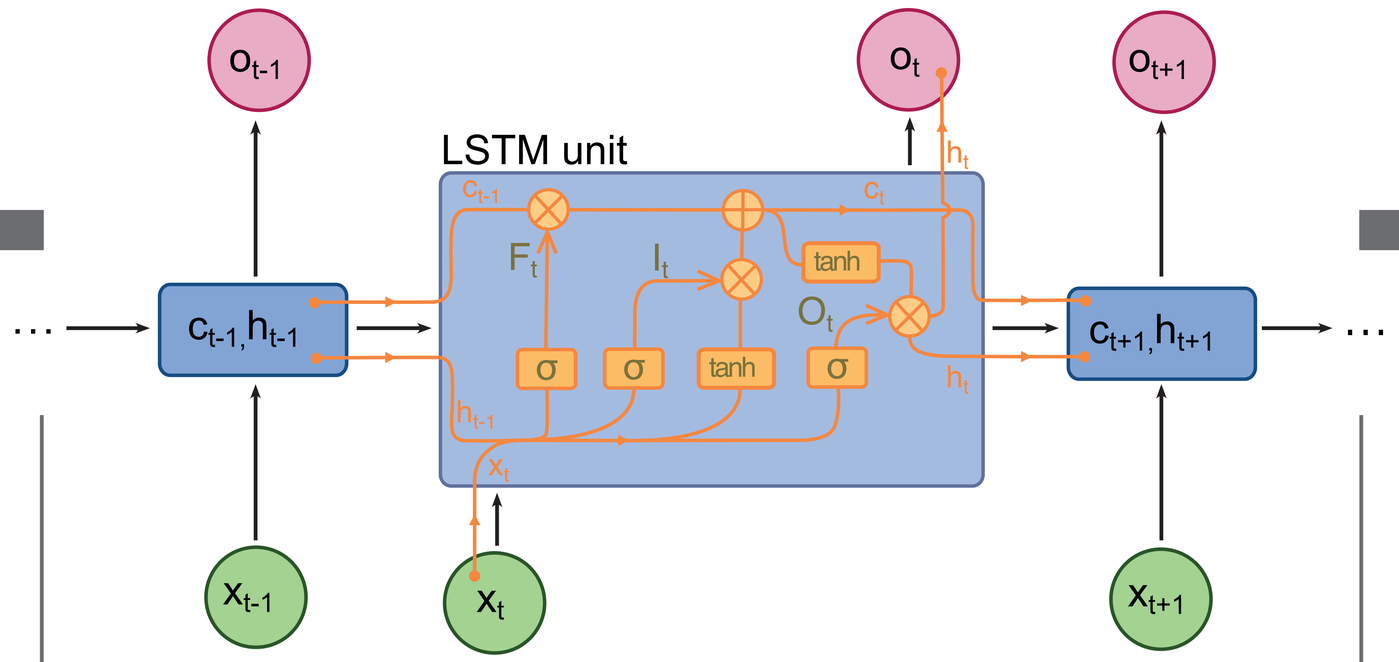
S Z U Æ.

Fraktur was among the most widely used fonts in Norwegian print up until the start of the 20th century. Spelling variation in the OCR-generated texts based on books from the period 1800-1850 is due both to OCR errors and to varying spelling conventions in the period.



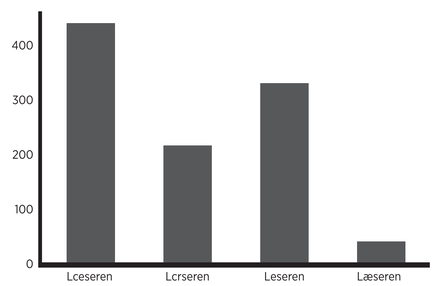
Annotating

In order to train neural networks for OCR, we have to rely on gold data or a ground truth i.e. data corrected or controlled by a human annotator. Tesseract, the system we have chosen as our baseline system, needs a scanned book page and a .box file that contains the textual content of the book with line relevant coordinates. To make this mapping from picture to text, we use the program jTessBoxEditor a sub-system of VietOCR.



Training

Our baseline neural network architecture is a Long Short Term Memory network (LSTM). These types of networks are specialized in predicting sequential data. When the dataset and model architectures are large enough, such neural networks demand high performance computing devices.



Evaluating

Quality of OCR can be measured with a number of metrics such as accuracy, error rate, precision, recall and F1-measure. We are using PRImA TextEval 1.4 to compute these metrics both on the character level and word level.