Improving the recognition of Dutch Gothic machine print, at four levels in the processing pipeline, in four days

Lambert Schomaker, Mahya Ameryan, Mirjam Cuper, Koen Dercksen, Jerry Guo, Rutger van Koert, Adriënne Mendrik, Konstantin Todorov, & Xue Wang

Report on an NWO/ICT with industry project (5-day sabbatical/hackathon) focusing on OCR problems at the Dutch National Library (KB) and the Huygens Institute, January 20-24, 2020, Lorentz Center, University of Leiden



1. Problem statement

Libraries and archives are struggling with optical character recognition (OCR) of old machine-print fonts such as Gothic or 'fraktur'. This font was used in many important historical printed collections such as administrative texts and the then (17th century) newly invented 'newspapers' with interesting and detailed reports on important developments and events. When applying current state of the art OCR tools or sending the scanned images to large well-known companies that provide OCR services, the returned results are still quite disappointing. Problems are observed at all levels in the processing pipeline: binarisation suffering from ink bleed-through, layout analysis suffering from deviating page designs, marginalia and graphics, character recognition suffering from lack of pertinent font examples and font variation (Roman/Gothic) in a document and, finally, linguistic post processing suffering from an utter lack of encoded digital text corpora of suitable size. Actually, the OCR process is often intended to arrive at such corpora in the first place.

2. Approach

A team was formed to approach these problems in four days, with a fifth day for reporting (other teams were working on other industrial problems at the Lorentz Center, this week). The team decided to address problems at all levels in the processing pipeline.

3. Team composition

Industry: Dutch National Library (KB) & KNAW Humanities Cluster (Huc)
- Mirjam Cuper and Rutger van Koert

Academia: - prof. dr. Lambert Schomaker (team leader, University of Groningen, Al/ML) In alphabetic order: Mahya Ameryan (Al, University of Groningen), Koen Dercksen (Radboud University), Jerry Guo (Technical university, Delft), Konstantin Todorov (ILLC, University of Amsterdam), dr. Adriënne Mendrik (e-Science center, The Netherlands), Xue Wang (CS,

Leiden University)

4. Tasks

The processing pipeline consists of (1) image preprocessing, (2) layout analysis and segmentation into meaningful text objects (lines, words, characters), (3) text recognition and (4) linguistic post processing.

Adriënne Mendrik focused on modeling the overall process in order to arrive at a specification of a modular work flow including performance evaluation, using her expertise on similar processing pipelines encountered at the Dutch e-Science center, in medical imaging applications.

Preprocessing: Jerry Guo (binarisation)

Layout analysis: Xue Wang (word segmentation)

Text recognition: Mahya Ameryan (characters), Lambert Schomaker (words)

Post processing: Koen Dercksen, Konstantin Todorov

5. Data & research questions

The first day of the week was used to delve into the problem. Schomaker gave a crash course at the whiteboard on OCR to the participants who were from different research fields and literature hints were disseminated. The problem owners presented a host of data sets and problems. The post-processing group discovered an SQL database with recognized text and wanted to relate it back to the original images. In itself, this appeared not to be possible because the work flow in the application domain is more focused on the encoded digital text than on image (pixel) related information. However, inspection of the Alto .xml files from the commercial OCR company that provided recognition results, it appeared that the bounding boxes (ROIs) of recognized strings were present in the .xml. This data set proved to be usable by all participants. It consists of a newspapers collection in Dutch Gothic script (1662 – 1795) from the Meertens Institute (Amsterdam), 15172 scans in Jpeg2000 format, which is common in libraries and archives, but not in science. The scans were converted to Jpeg by van Koert, who also provided a server with GPUs. Collaboration tools were Google Drive, Slack and Surf storage. The ensuing research questions (RQ) are:

- 1. Can we improve the recognition results by improving the binarisation of the scans?
- 2. Can we improve the detection of words, because the commercial OCR providers often produce very long strings in the recognized output, with no blanks between words
- 3. Given the problems at the character level, wouldn't it be better to train at the level of words, and how fast can we collect word labels from a tabula rasa setting, using high-performance computing and a human in the loop (the Monk approach)
- 4. Contrarily, can we use the suboptimal recognition results from a commercial provider (i.e., without human-validated ground truth) to train our own recognizers from scratch and allow many recognizers to be applied to the same data?
- 5. What are the effects of using state-of-the art text transduction tools to improve the quality of the OCR output to something that is closer to the intended language?
- 6. How does the performance evaluation work, at the different four stages in the processing pipeline?

6. Methods

RQ1. Binarisation – The approach is to apply a number of traditional and deep-learned binarisation methods to a selected subset of scans from the Meertens newspaper data. Because no ground truth on the pixel intensities (ink/background) exist, as in the international DIBCO competition, another performance indicator is used, i.e., the final recognition performance of a common free OCR tool, Tesseract, on the data, which is used as the measuring device.

RQ2. Word segmentation – The HUC (problem owner) already had experience with using ARU-net (a U-net CNN, i.e. a deep-learning method) for baseline estimation. Because the Alto .xml output contains the ROI on the position of blanks via the '<SP>' tag, a new network can be trained, end to end, combining the line segmentation and word segmentation in an innovative single pixel-to-pixel transform.

RQ3. Full-word recognition – The existing Monk e-Science service was used to bootstrap a label collection process, starting from the tabula rasa condition), i.e., no labels. Scans were segmented into lines automatically and oversegmented into word candidates. Users start to label some correctly segmented words and an immediate process of data mining starts in the background using nearest-neighbour matching (upon the first label) and nearest-mean matching (with Nlabels > 1), in a high-performance computing setup. This is done using an alternation between recognition and retrieval, yielding ranked hit lists that can easily be labeled: the 'Fahrkunst' principle, Schomaker & van Oosten (2014).

RQ4. Fresh, end-to-end training of a CNN/LSTM using the suboptimal recognition results of a commercial OCR provider as the training set. The OCR output is pruned to contain word-like strings, starting with a letter. A small ensemble of five classifiers using plurality voting and tie resolution is used. The performance evaluation is case insensitive.

RQ5. A current state-of-the art approach to language translation is to use word embeddings (cf. Latent semantic indexing and the later word2vec approach) followed by an LSTM encoder and an LSTM decoder. For word embeddings, we use BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018), a pretrained network that is fine tuned on the text to be expected in the Meertens Newspapers corpus.

RQ6. How to provide a modular pipeline and work flow, such that users in the application domain can quickly evaluate different variants of processing? Ideally, tests are performed on the grounds of a neutral party, such as a national e-Science server. Is this possible in all cases, e.g. using virtualization or 'Docker' containers? In any case the overall concept for the framework can be designed in the short time frame of the current project.

7. Results

RQ1 – binarisation

Jerry Guo demonstrated that binarisation methods are less important than envisaged. The traditional methods deteriorate the OCR results. Deep-learning (Deep Otsu) does not deteriorate the recognition performance noticeably, but leaving the input images – as is – gives the OCR performance that can be reached. Tesseract was used as the 'thermometer'.

RQ2 – word segmentation

Xue Wang showed that deep-learning based line segmentation (ARU-net based) can be enhanced by adding word-segmentation suggestions, in a pix2pix manner. The model was trained end-to-end, using recognition results (Alto.xml) of a commercial provider.

RQ3 – full-word recognition, from tabula rasa

Lambert Schomaker and Mirjam Cuper showed that the Gothic text could be ingested by the Monk system. Manual labeling with machine support yielded 319 word classes, 1760 human-labeled images in under two hours. This data can be used to evaluate different recognizers.

RQ4 – training a new recognizer on the basis of other OCR

Mahya Ameryan showed that an LSTM developed at Al@RuG could be trained end-to-end, achieving a word-recognition rate of 88%. Notably, the recognizer was trained on the output of a commercial recognizer, i.e., using word regions of interest and their OCR text as ground truth (100% can not be reached). The result is favorable and compares to a character recognition accuracy of 97.5% for a language with 5-letter words, on average.

RQ5 – linguistic post processing

Koen Dercksen and Konstantin Todorov used the famous BERT (word-embedding method by Google) system to fine tune it (post train it) using the Meertens Newspaper texts, with promising results.

RQ6 – performance-evaluation framework

Adriënne Mendrik modeled the work flow and processing pipeline in order to realize a modular, adaptable framework in which individual component variants can be inserted to assess their effect on evaluation. Performance metrics are considered within their particular usage context, e.g., transcription and retrieval require different performance metrics, and the object under consideration (characters vs words) also plays a central role in understanding the performance of an OCR system.

8. Conclusion

This was an exciting project week with tangible results, that are expected to have an impact on the 'industry', i.e., the KB and Humanities Cluster (HuC).

Notable is the fact that there is much more information in the recognition results of commercial providers than is currently used. Massive amounts of training data can be extracted. Apparently good recognition results can be achieved by alternative recognizers that

are trained on such imperfect data. By using user-friendly labeling tools, benchmark data sets can be developed by the institutes themselves, which puts them in a more comfortable position in dealing with OCR providers. The results present a promising picture regarding the future integration of Gothic-font documents in current search engines such as 'Delpher'.

9. References

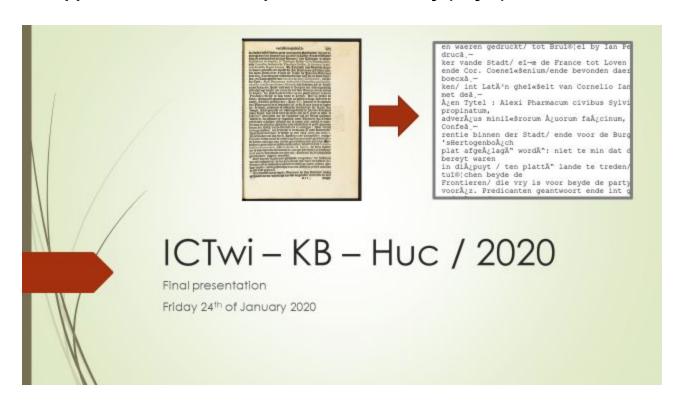
Ameryan, M. & Schomaker, L.R.B. (2019), A limited-size ensemble of homogeneous CNN/LSTMs for high-performance word classification, arXiv:1912.03223

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805 [cs.CL]

Gruening, T., Leifert, G., Strauß, T., Michael, J. & Labahn, R. (2019). A Two-Stage Method for Text Line Detection in Historical Documents, arXiv:1802.03345 [cs.CV]

van Oosten, J.-P. & Schomaker, L.R.B. (2014). Separability versus prototypicality in handwritten word-image retrieval, Pattern Recognition, 47(3), pp. 1031-1038

10. Appendix – Final slides presented on Friday (day 5)





KB-Team (only the participants present at Lorentz Center during the week)

- Industry: KB & Huc KNAW
 - Mirjam Cuper (performance evaluation)
 - Rutger van Koert (tool benches image & text processing)
- Academia (team leader: prof. dr. Lambert Schomaker)
 - Mahya Ameryan (Al, RUG) text recognition
 - Koen Dercksen (Radboud Univ.) multimodal information processing
 - Jerry Guo (TU Delft) computer vision
 - Konstantin Todorov (ILLC, UvA) computational linguistics
 - dr. Adriënne Mendrik (e-Science center) performance evaluation frameworks
 - Xue Wang (CS, Leiden Univ.) object vision

KB What is the problem?

Mirjam Cuper Rutger van Koert

en waeren gedzuckt/ tot Bzuffel by Jan De her bande Stadt/ en de France tot Loben ende Coz. Coeneftentum/ende bebonden daer hen/ int Latijngfeftelt ban Cornelio Ianfen fen Tptel: Alexi Pharmacum civibus Sylvi adversus ministrorum suorum fascinum, end rentie binnen der Stadt/ ende booz de Burg plat afgef lage wozde: niet te min dat defe twe in dispupt / ten platte lande te treben/ op eer Frontieren/ die byp is boog bepbe be part boogly. Debitanten geantwoogt ende int geh en waeren gedruckt/ tot Brui@;el by Ian Pe drucă,-ker vande Stadt/ ei-s de France tot Loven ende Cor. Coeneixăenium/ende bevonden daer

ken/ int Lată'n ghei«šelt van Cornelio Iar

met deă -Ajen Tytel : Alexi Pharmacum civibus Sylvi

propinatum, adverājus miniteārorum Ājuorum faĀjcinum, Confeā, rentie binnen der Stadt/ ende voor de Burg

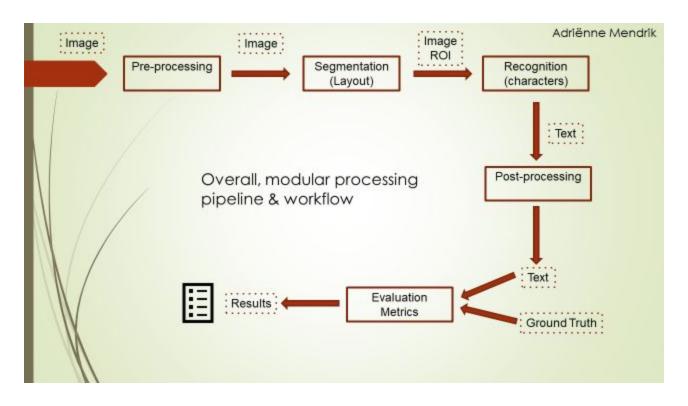
rentie binnen der Stadt, ende voor de Burg 'sBertogenboäch, plat afgeäkjlagä" wordä": niet te nin dat d bereyt waren in diäjpuyt / ten plattä" lande te treden/ tul®jchen beyde de Frontieren/ die vry is voor beyde de party vooräjz. Predicanten geantwoort ende int g

KB) national library of the netherlands

KNAW Humanities Cluster

Delpher



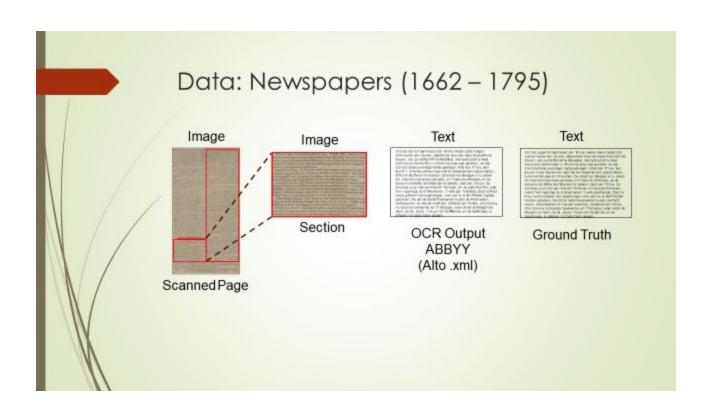


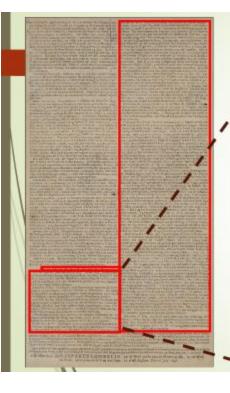
Data & environment

- Among all the data, we found a set that suited several sub teams: the Meerten's set. Actually Koen wanted to go from plain text back into the Alto XML, which was not possible, but then it appeared that ABBYY put the bounding boxes of recognized fragments in this metadata
- Usable by Xue, Mahya, Lambert, Koen, Konstantin, Jerry (i.e.all), in each of the processing stages
- Conversion needed: Jpeg2000 (archives & libraries) is not often used in science. It needed to be converted to regular .jpg
- Rutger provided server access with GPUs, and big external disks with data
- We used: Google Drive, Slack, Surf storage
 Tensorflow & own tools, as well as some new coding (Python, Bash, C)

Alto.xml format snippet (Meertens collection): labels down to the word ROI in the page image

```
<TextBlock ID="P1 TB00003" WIDTH="1409" HEIGHT="100" VPOS="253" HPOS="112"</p>
           STYLEREFS='TXT 2 PAR LEFT'>
   <TextLine ID="P1_TL00003" WIDTH="405" HEIGHT="2" VPOS="252" HPOS="112">
     <String ID="P1 ST00004" WIDTH="43" HEIGHT="2" VPOS="252" HPOS="112"</p>
         CC="000000000" WC="0.96" CONTENT="Utrechtfe"/>
     <SP ID="P1 SP00002" WIDTH="18" VPOS="254" HPOS="155"/>
     <String ID="P1_ST00005" WIDTH="69" HEIGHT="2" VPOS="252" HPOS="172"</p>
         CC="0000" WC="0.96" CONTENT="Vry-"/>
     <SP ID="P1_SP00003" WIDTH="18" VPOS="254" HPOS="241"/>
     <String ID="P1 $T00006" WIDTH="103" HEIGHT="2" VPOS="252" HPOS="259"</p>
         CC="000000" WC="1.00" CONTENT="daegfe"/>
    <SP ID="P1_SP00004" WIDTH="18" VPOS="254" HPOS="361"/>
     <String ID="P1_ST00007" WIDTH="138" HEIGHT="2" VPOS="252" HPOS="379"</p>
         CC="00677350" WC="1.00" CONTENT="Courant."/>
   </TextLine>
</TextBlock>
```





Original Scan: Example Segments

LEDEK LANDEN.

Tyt het leger tot Gembloers den 16 Iuly. Heden dele morgen arrivertde weder een Coutier, afgelonden door den Heer Keutvoist van Saxen, aen zyn Brittanisehe Majesteyt, met tyding dat hy Heer Keutvoist de Franssen in d'Arrier-Guarde was gevallen, en een Considerabele avantagie hadde gekregen.

Aelft den 17 Inly. Syn Excell: onse Gouverneut legt met sijn Detachement nog tot Haren, tusselnen Brussel en Vilvoorden. De Graef van Bergeyk is nu alleen tot Intendant Generael gemaekt, om 't Geld te ontrangen, en de Spaensche Militie alle Maenden te betalen.

Gent den 18 Iuly. Op Sondag quam hier een Fransch Tamboer, om de party Franssen, laetst hier ingebragt, te rantsoeneeren; 't welk gechiedde. Doch la Fleur wierd gisteren hier opgelangen, voor wien fy al 300 Pistolen hadden geboden, Na dat de Gerantsoeneerden buyten de Poort waren, deserreerden 'er met der haest tien.

Oostende den 18 July. Ons Convoy na Spanjen bestaende urt 17 Schepen, waer onder de Maegd van Gent, de St. Jacob, 't Huys van Oostenryk, en de Gastanega, is gisteren na Cadix t'zeyl gegaen.

Bruffel den 18 Iuly. Sondag trokken 2 Brandenburgte Regimenten te voet na 't Campement van den Marquis de Gaffanaga,dat nu maer ten uur van dele Stad leyt , en Macndag 't Regiment Dregonders van

Example ABBYY OCR Output

eerde weder een Courier, afgefonden door den Heer Keurvorft van Dien, aen zyn Brittanifebe Majesteyt, met tyding dat by Heer Keur-orft de Fransfen in d'Arrier-Guarde was gevallen, en een Conside-

bele avantagie hadde gekregen. Aeift den 17 Inly. Syn Excelli onfe Gouverneur legt met fijn Detamenent nog tot Haren, tuffchen Bruffel en Vilvoorden. De Graef an Beeggyk is nu alleen tot Intendant Generael gemackt, om 't Geid contiangen, en de Spieniche Militie alle Maenden te betalen. Gest den 18 Iuly. Op Sondag quam hier een Franch Tamboer, om

te party Frantien, iserit hier ingebragt, te rantioenceren; 't welk ge-chiedde. Doch la Fleur wierd gifteren hier opgehangen, voor wien val : Pistolen hadden geboden, Na dat de Garantsoeneerden buy Poner waren, deserreerden 'er met der haeft rien.

Orflende den 18 laty. Ons Gonvoy na Spanien bestaende uyt 17 Chepen, waer onder de Mased van Gent, de St. Jacob, it Huys van Jostenryk, en de Gastanaga, is gisteren na Codixt zeyl gegaen. Bonstei den 18 laty. Sondag trokken a Brandenburgte Regimenten

et na 't Campement van den Marquis de Gaffanaga,dat nu moer n unt van dele Stad leyt, en Maendag 't Regiment Dregonders van

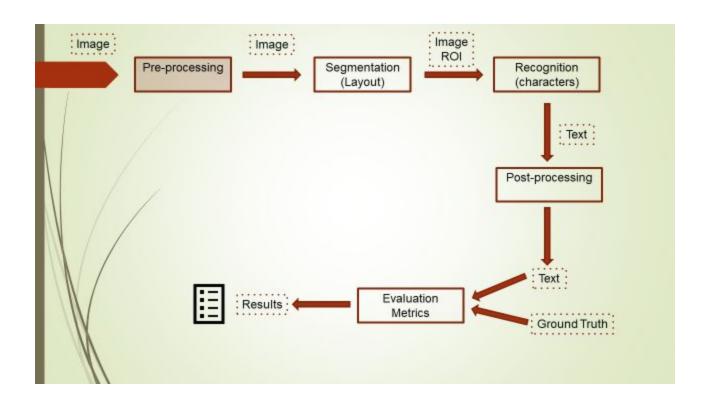
Vyt het Irg'r tot Gernhloers den 16 Inly. Heden defe morgen arrlâ weder een Courier, afgefonden door den Heer Kcutvotft van Saxen, acn zyn BrittanifA"he Majefteyt, met tyding dat hy Heer Keürvoitl de Franffcn in d'Arricr-Guarde was gevallen en een Conlidcrabele avantagichadde gekregen. Ar!]} den 17 luly. Syn Excă"ll: onfe Gouverneur legt met fijn Detachement nog tot Haren, ttiffclisn Bruffel en Vilvoorden. De 6raef| van Bergeyk is nu alleen tot Intendant Generael gemaekt, om 't Geld te ontfangea, en de Spaeniche Militie alle Maenden te betalen. Geit den 1S luly. Op Sondag quain hier een Fianfch Tamboer, om de party Franffcn, betil hier ingebragt, te rantfoeneeren; 't welk ge- ' fchiedde. Doch la Fleur wierd gifteren hier opgehangen, voor uien fy al 30-Piftolen hadden geboden, Na dat de Gerantfoeneerden buyton de Pooit waren, deferteerden 'et met der liaeft tien. Oofte.tdi Jen 18 lttly. ()ns Convoy na Spanjen bellaende uyt 17 Schepen, waer onder de Maegd van Gent, de St. Jacob , 't Huys van OoA¶enryk, en de Gaftanaga, is gifteren na Cadix t'zeyl gegaen.

ABBYY OCR Output

Vyt het Irg'r tot Gernhloers den 16 Inly. Heden defe morgen arriå weder een Courier, afgefonden door den Heer Kcutvotft van Saxen, acn zyn BrittanifÄ he Majefteyt, met tyding dat hy Heer KeÅ1/4rvoitl de Franffcn in d'Arricr-Guarde was gevallen , en een Conlidcrabele avantagichadde gekregen. Ar!]} den 17 luly. Syn ExcA"ll: onfe Gouverneur legt met fijn Detachement nog tot Haren , ttiffclisn Bruffel en Vilvoorden. De 6raefl van Bergeyk is nu alleen tot_Intendant Gcnerael gemaekt, om 't Geld te ontfangea, en de Spaeniche Militie alle Maenden te betalen. Geit den 1S luly. Op Sondag quain hier een Fianfch Tamboer, om de party Franffcn, betil hier ingebragt, te rantfoeneeren; 't welk ge- ' fchiedde. Doch la Fleur wierd gifteren hier opgehangen, voor uien fy al 30- Piftolen hadden geboden, Na dat de Gerantfoeneerden buyton de Pooit waren, deferteerden 'et met der liaeft tien. Oofte tdi Jen 18 Ittly. ()ns Convoy na Spanjen bellaende uyt 17 Schepen, waer onder de Maegd van Gent, de St. Jacob, 't Huys van OoA¶enryk, en de Gaftanaga, is gifteren na Cadix t'zeyl gegaen.

Ground Truth

Uyt het Leger tot Gembloers den 16 luly. Heden dese morgen arri veerde weder een Courier, afgesonden door den Heer Keurvorst van Saxen, aen zyn Brittanische Majesteyt, met tyding dat hy Heer Keurvorst de Franssen in d'Arrier-Guarde was gevallen, en een Considerabele avantagie hadde gekregen. Aalst den 17 luly. Syn Excell: onse Gouverneur legt met sijn Detachement nog tot Haren, tusschen Brussel en Vilvoorden. De Graef van Bergeyk is nu alleen tot Intendant Generael gemaekt, om 't Geld te ontfangen, en de Spaensche Militie alle Maenden te betalen. Gent den 18 luly. Op Sondag quam hier een Fransch Tamboer, om de party Franssen, laetst hier ingebragt, te rantsoeneeren; 't welk geschiedde. Doch la Fleur wierd gisteren hier opgehangen, voor wien sy al 300 Pistolen hadden geboden, Na dat de Gerantsoeneerden buyten de Poort waren, deserteerden 'er met der haest tien. Oostende den 18 luly. Ons Convoy na Spanjen bestaende uyt 17 Schepen, waer onder de Maegd van Gent, de St. Jacob, 't Huys van Oostenryk, en de Gastanaga, is gisteren na Cadix t'zeyl gegaen.





Bottom-up word segmentation using a pixel to pixel deep-learning tool

Xue Wang

together with Rutger & Lambert

Xue Wang

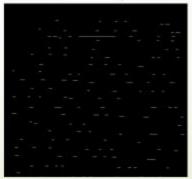
- Goal: Model --->find the word position & spaces in the image
- Dataset: image & alto.xml (with annotation about spaces and text strings)
- Step 1: clean the data and make the ground truth

Raw image

V R A N C K R Y C K.

Aris, des 11 Angelis. Den Hettog van Sunderlant gontnumeers hier in belogens; doen
de groote klachten over onie Fregurien, die
hubben gewegers tei frigiens voor fijn Kongemaar vint weynig gehoorgalssook over de foltie voor den Hettog van Lotteringen, wans fijn
theyt oosdeelende van dat Høys te veel fehade
from geleden se hebben, wil buyten de prefenniees doen in fijn voordeel. Den Heese Grahal doer in feilicks groose devoir tot de evacuade de Steden, water in fijne Mairfleyt ook vooen heeft; doen onder voorwactden van 1 van
oormaneelt Placefes in jin vygen bewaring te
en, tot'er eij dat Brandenburg achnortliche fien stot'er eij dat Brandenburg achnortliche fien of verden fol preserven hebben. De see

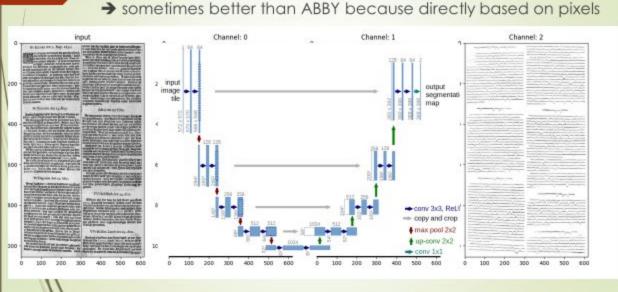
Ground-truth:space

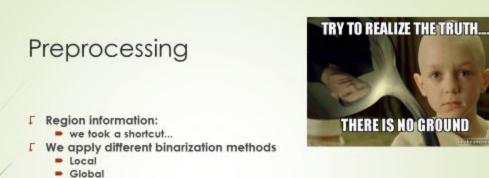


Ground-truth:string



Train Model: based on ARU-Net (output: the word and spaces in the image!)





Both on plain grayscale input and DNN enhanced documents

We don't have ground truth for binarization We decide to compare the OCR results directly

Tesseract

Preprocessing

For the trees set Goodsteer des 16 July. Heen dele morgen enversids weder een Countier, afsectonden door des Heer Kourvoord van Saxen, aan zur Brittsmidde Majedleye, met triding dat by Heer Kourvoord de Frantien in d'Arriet Gorarde was gevallen, en een Countiertable is vantagge halde gekregen.

As/8 des 19 faty. Syn Excell: onfe Gouverneur legt mer fijn Dera chrimant ung tor Haren, tiffichen Beufiel en Vilvoorden. De Grei van Bergeyk is nu alben toe Internatue Geinensel gemacht, om 's Geld to omthogen, en de Spaeniche Militie alle Maeuden te bezalen.

de party Francien, lierel hier ingebrage, se rantioenceren 1 t well gi fehreide. Doch la Pleat wierd gisteren hier upgelingen, voor we fr al -- Pitolen ladden gehoden, Na dat de Gerantioencerden im tinde Poor waren, del creenion er mer der hard frien.

Schepen, weet onder de Mingel van Gent , de St. Jacob , 't Hay's van Oodhenysk, en de Gaffamega, is guiteren na Gaffart zoyl gegaen. Broden den 18 Iniv. Sondag trokken a Brondenburgie Regintensen

Plain input

Pyr her leger in Grenhister deu 15 telp. Heden dele inonym sei ver in weder een Comret, afresloden door den Heer Eurevorft va Sorm, noer syn Burantische Majester, am 194mg dat hij blees Kevord de Frantisch in d'Arriert Georde was gevallen, en een Comitor

As if den 19 faty. Syn Excell onfo Government legt merfinn Den chemont ong tax libres, reficient finific en Vilvoorden. De Gele van bergryk in nu allien tot Intendant General gemacht, one 'i Gol te omfangen, en de Syneniche Militie alle Maenden te betalen.

de party Frantien, best hier regionagt, se ranticenceren, i verils ge ichiedde. Duch la Pleur wierd gideren hier opgehangen, voor wie fy al pro Piñolese hadden geboden, No dat de Geraumforneerden huj han de Poorty wagen, delgereerdom 'et met det hard tiese.

Company dea 18 Inty. One Control in a spanier betterning up of Schepen, where conder the Mangel van Gent, de St. Jacob., 't Huys an Owlewayk, en de Galtanega, is giberen us Godin e zeyl segaen.

Brade dea 18 Inty. Sondag trakken a Brandenburgie Regimenter.

Brahet des 18 July, Sondag trokken 3 Erandenburgle Regimenter te vort ma 3 Campement van den Marquis de Gaffonstagdet nu met sen unt van dele Sand Fret, en Marculos 3 Regiment Desconders van

Enhanced by DeepOtsu

Preprocessing

Plain input + adaptive mean Tys her Leger ist Gembleres den 16 July. Heden dele morgen strivectes wedet een Couriet, afgesonden door den Heer Keurvorst van Saxas, aan zyn Brittanische Majesteyte, met tyding dat hy Neer Kourvoust de Fransisen in d'Arriet-Gaarde was geralten, en een Considerabele avantagie hadde gekregen.

Anfil den 19 July. Syn Excell: onste Gouverneut legt met sijn Detachennen en grot Haren, unstehen Innsiste en Visvoorden. De Gezef van Bergeyk is nu alleen sot Intendant Generael gemackt, om 't Geld te onstangen, en de Spasensiche Militie alle Manedien te beralen. Gest dies 18 July. Op Sondag quam hier een Fransich Tamboer, om de parsy Fransisen, besteh heit ingeforgt, et ransieenceten; 't welf geschiedde. Doch la Fleur wierd gisteren hier opgelangen, voor went fy al pop Fisholen hadden gebuchen. Ná dat de Geransforneerden buyten de Poort waren, descreterden 'et met de hart fitten.

Onlinede den 18 July. Ons Gowoy on Spanjen bestande uye 17 Schegeo, were ondet de Mazegt van Gent, de St. Juscho, 't Huys van Oudenryk, en de Gallanega, is gisteen na Cader Tayl gegaen.

Bristil den 18 July. Sondag trokken a Brandenburge Regimenten te vort na 't Campement van den Mazeguis de Castanaga, dat no more een uut van dese Stad leyt, en Macndag 't Regiment Dregoodetz van

Enhanced by DeepOtsu + adaptive mean

Preprocessing

"Up het Lege est Gemètere der 16 July. Heden dele morgen ariv vertek weiter ein Courier, afgefonden door den Heer Reutvorst van Saxen, am zyn Britanische Majeffere, met tyding dar hy Neer keur vout de Frantien in d'Arrier Gonade was gevillen; en een Condetthable krantagie halde gebregen.

And den 17 July. Syn Excell onde Gonzeneur legt mee fijn Denrichment nog tot Haren, enflichen Vilvoorden. De Graef van Bergeyk is nu alben tot Intendant Gonerael gemacht, om 't Geld yo ontengen, en de Spariniche Militei alle Maneden te bestim.

Grae den 18 July. Op Sondag quam hier een Franch Tamboer, om de patty Ernaffen, besti hier ingebraget, te raniformeren; i veul gefehrdde. Doch is Skent wierd gifteren hier opgelangen, voor wom 'y al 300 Pistolen hadden gehoden. Na dar de Gerantioeneerslein beijsten de Poort waren, delerteerden et met det haeft van 18 July. Goes Connoy na Spanien beltende urg 177. Schepen, West word onder de Mazed van Gont, de St. Jacob ; t Hysylvan Ousteneyk; en de Gistanoga, is gisteren na Cadiet veri ergaen.

July dar 4 Stair. Sondag rooken a Brandenburge Legementen te vort na 11. Campement Van de Marquis de Gallanoga/da nu maer een uur van deste stad leyt, en Maendag it Regimen Dragonders van

Plain input + adaptive gaussian Hys hes Leger ses Gembieres den 16 16/19. Heden dele morgen setivectée wedet een Couriet, afgelonden door den Heer Keurvorft van
Sozan, aan zyn Brittanilehe Majefleyt , met tyding dat hy Neer Keurvorft de Frantien in d'Atriet-Guarde was gerallen , eo een Confiderabele avantez je hadde gekregen.

Anfil den 19 14/2. Syn Excell: onle Gouverneux legt met fijn Deschennent oop tot Haren, vuiffelm Bruffle en Vilvoorden. De Gezef
van Bergeyk is nu alleen tot Intendant Generael gemackt, om 't Geld
te ontiangen, en de Spacenliche Militie alle Manedton te beralen.
Gesa den 18 16/2. Op Sondag quam hier een Frantich Tamboet, om
de parsy Frantien, bezeit hiet ingefroatge, te rantelonerten j't veilt gefehiedde. Doch la Fleur wired grifteren hier oppelangen, voor wen
fy al 100 Pilholen hadden geboden. Na dat de Getamtforneerden buyten de Poort waren, deletzeerden 'et met der harft tien.

Oulfrede den 18 16/2. Ons Gonoop na Spanjen beflaende urg 17
Schegen, west ondet de Mazeri van Gent, de St. Jacob, 't Hujs van
Oulfrenyk, en de Gultanega, is gifteren na Cadert zeyl gegaen.

Brafiel der 18 16/19. Ons Gondag trokken a Brantehustef Regimenten
te voet na 't Campement van den Mazerius de Caffanaga, dat no more
een uut van dele Stad leyt, en Macndag 't Regiment Dregooderts van

Enhanced by DeepOtsu + adaptive gaussian

Preprocessing

Hys hes Leges tot Genkleres den 16 July. Heden dele morgen artiverede wedet een Couries, afgefonden door den Heer Keurwork van
Saxen, aan zen Britannische Majesteye, mee tyding dat hy Neer Keurvoust de Frantien in d'Artier-Guarde was gevallen, en een Considertabele zwantagie landde gekregen.

As if den 19 July. Syn Exectlit onde Gouverneur legt met fijn Detachement ong tot Haren, tutticken Beutlid en Vilvoorden. De Graefvan bergeyk is nu alleen tot Intendant Generael gemacht, om 's Geld
te omtfangen, en de Spacentiche Militie alle Maneden et beralen.

Gest den 48 July. Op Sondag quam hier een Frantich Tamboer, om
de pasty Frantien. Jetekt hier ingebragt, te rantionereren; 't welk gefehielde. Doch la bleur wired githeren hiet opgehangen, voor wenfra 1; - Pittolen hadden gehoden. Na dat de Gerantitoenereden buyten de Poort waren, deletteerden 'et met det Aardt tien.

Osflends den 18 July. Ons Gemory in Spanien betkande upt 19
Schegen, wet onder de Mazest van Gent, de St. Jacob. 't Hugs van
Osflenerek, en de Gallanega, is giferen in Coafet vzeyl gegen.

Brafiel des 18 July. Chas Gonger in Brandenburgte Regimenten
te vort en 't Causpement van den Mazendag 't Regiment Dragondett van

Plain input + Niblack

Tys het Leger ist Gembleres den 16 belg. Heden dele morgen artiverede wedet een Couriet, afresonden door den Heer Keurvorft van
Saxen, aan zyn Brittanische Majesteyte, met tyding dat hy Neer Keurvort de Frantien in d'Arrier-Genatde was gerakten, en een Considertabele avansige hadde gekregen.

Anfil den 19 lafe. Syn Excell: onste Gouverneur legt met fijn Detstennent oog not Heren, tussfeiten Brussle en Vilvoorden. De Gezef
van Bergeyk is nu alleen tot Intendant Generael gemackt, om 't Geld
te onstaingen, en de Synsensiche Militie alle Maneden te beralen.
Gest des 18 lafe. Op Sondag quam hier een Fransch Tamboer, om
de parvy Franssen. Het ingeforigt, te rantscenceren; 't welk gescheidede. Doch la Fleur wierd gisteren hier opgelangen, voor wier
fy al 300 Pistolen hadden geboulen. Na dat de Geramsfonneerden bayten de Foort waren, descreerden 'et met det harst tiene.

Onstenda den 18 lafe. One Gonvoy na Spanjen bestaande uyt 17
Schegen, water onder de Mazegi van Gent, de St. Jacob, 't Huys van
Oustenryk, en de Guitanega, is gisteren na Caden 't zeyl gegaen.

Brassi der 18 lafe. Sondag trokken a Brandenburge is Regimenten
te vost na 't Campement van den Mazeguis de Castanaga, du nu mer
een uut van dese Stad leyt, en Macredag 't Regiment Dragonderz van

Enhanced by DeepOtsu + Niblack

Preprocessing

Fyr het leger tet Gembleres den 16 lady. Heden delle morgen artiwertig werder ein Courier, afresonden door den Heer Keurworft van
Sown, aan zen Brittanische Majestere, met tyding dat hy Heer Keurwordt de Frantien in d'Arrier-Guarde was gevallen, en een Courier
sabel e vantagie hadde gekregen.

Arif den 17 lady. Syn Exsellt onfe Gouverneur legt met fijn Detachement ong tot Haren, untildum Benfiel en Vilvoorden. De Graefvan Bergryk is nu alben tot Intendane Generael gemackt, om 't Geld
van Bergryk is nu alben tot Intendane Generael gemackt, om 't Geld
van ontringen, en de Spoenische Militeit alle Maenden te bershen.

Grad den 48 lady. Op Sondag quam hier een Frantich Tamboer, om
de passy Frankfan, best hier ingebraget, te ranticureren; 1' well gefehreide. Doch la vieuw wierd gifteren hiet opgehangen, voor wom
fy al ; - Piftolen hadden gehoden, Na dat de Gerantboneerden bayten de Poort waren, debriteerden 'et thet det haeft rien.
Opfrode den 18 lady. Onse Gonnoy na Spanien betlaende upt 17
Schregen, west onfert de Muzgel van Gent, de St. Jacob , 't Hary van
Ouffreet k, en de Gallanoga, is gifteren na Godix i zogl gegen.

Bradei det 18 lady. Sondag trokken i Brandeinburge Regimensen
te wort na 't Gauspement varsifen Marquis de Gaftanoga, ian nu meer
cen mat van dele Stad leyt, en Maendag 't Regiment Dragondett van

Plain input + Sauvola

Thy her leger tot Grahlerst den 16 lely. Heden dele morgen striverede wedet een Courier, afredonden door den Heer Kouworft van Soven, aan zwa Brittaniiche Majedheye, met tyding dat hy Heer Kouwordt de Frantien in d'Arrier-Gausde was geralen, en een Cominterale, le avantagie hadde gekregen.

Ar fil den 19 lafy. Syn Eucellt onde Gouverneur lege met fijn Dear-Cement ong tot Haren, untificium Bruffich en Vilvoorden. De Greef van bergryk is nu alleen tot Intendant Generael gemackt, om 't Geld te ontfangen, en de Synsenfiche Militie alle Manedone te beralen.

Grad den 68 lafy. Op Sondag quam hier een Frantich Tamboer, om de parsy Frantien, beet hier ingebragt, et rantienerenen; 't well gefehendde. Doch la Heur wired gaiteren hier opgehangen, woor wird fy al gan Pitholen hadden geboden. Na dat de Gertamformeerden huyten de Poort waren, deletteenden 'er met der hardt itee.

Oulfrede den 18 lafy. Om Gonovo na Spanien beflaende upr 17 Schapen, weter onder de Mazed van Gent, de St. Jacob, 't Huys v.

Oudtroyk, en de Galtanega, is gifteren na Gadet veryl gegoen.

Bradfel der 48 lafy. Sondag trokken a Brandenburgte Regimenten te wort na 't Campement van den Mazedwy't Regimen Dregondert van

Enhanced by DeepOtsu + Sauvola

Preprocessing

Flys het leger tot Gembleres den 16 ledy. Heden dele morgen artiwerde welet een Courier, afgefonden door den Heer Kenrvorst van
Saxen, aan zon Brittanische Majesteyt, met tyding dat hy Heer Keurword de Fransien in d'Arrière-Guarde was gevallen, en een Cousider
rabe le vrantagie halde gekregen.

Ae's den 17 ledy. Syn Excelli onse Gouverneur legt met fijn Detachement ong tot Heren, tussichen Beussel, om 't Geld
van bergeyk is nu albeen toe Intendant Gouernel gemackt, om 't Geld
van bergeyk is nu albeen toe Intendant Gouernel gemackt, om 't Geld
van bergeyk is nu albeen toe Intendant Gouernel gemackt, om 't Geld
van omfangen, en de Spaensiche Milliei alle Maenden te betalen.
Gest den 18 lety. Op Sondag quam hier een Fransich Tumboert, om
de party Fransien, lesett hier ingebragt, te ransificeneren 1, 'welk geschiedde. Doch is siem wired gisteren hier opgelungen, voor worn
sy's 3 2 Pistolon hadden geboden, Ni dat de Gerantioneren 1, 'welk geschiedde den 18 leit. Ons Convoy in Spanien bestande upt in
Schegen, west onder de Mangel van Gent, de St. Jacob, 't Hurs van
Oudreevk, en de Gustanega, is gisteren in Cadix 'zeg's gegaen.
Brasie des 18 leit. Sondag trokken is Brandenbargte Regimenten
te voor na 't Campement van den Marquis de Gastanaga, den na meer
cen mus van dese Stad leyt, en Maender 2 Regiment Dragondets van

Plain input + Otsu

Tys het Leger ist Gembieres den 16 July. Heden dele morgen artiverede wedet een Couriet, afgefonden door den Heer Keurvorft van
Saxan, aan zyn Brittaniiche Majefleyt , met tyding dat hy Heer Keurvortt de Franties in d'Arriet-Gaarde was geralkin , en een Confidertabele avange in halde gekregen.

Anfil den 17 July. Syn Excell: onde Gouverneut legt met fijn Dentement oog net Heren, tuffelem Bruffie en Vilvoorden. De Greef
van Bergeyk is nu alleen tot Intendant Generael gemackt, om 't Geld
te ontfangen, en de Syacenfiche Militie alle Manedien te beralen.
Gest des 18 July. Op Sondag quam hier een Franfich Tamboer, om
de parvy Franfien, beet hier ingeforigt, et ranticenceren; 't welk gefehiedde. Doch la Fleur wierd gifleren hier opgebangen, voor wier
fy al op Pitholen haden geboden. Na dat de Getannfonneerden bayten de Foort waren, deletzeerden 'et met det harft tien.

Onfleuds den 18 July. Ons Gonvoy en Spanjen beflaende uyr 17
Schegen, water onder de Mazgi van Gent, de St. Juscho, 't Huys vaOudensyk, en de Gultanega, is gifteren na Cadiat 'zeyl gegaen.

Braftel des 18 July. Sondag trokken a Brandenburgle Regimenten
te vost en 't Campement van den Mazquis de Caffanaga, dat nu more
een uut van defe Stad leyt , en Macndag 't Regiment Dregoodert van

Enhanced by DeepOtsu + Otsu

Preprocessing

Fire het leger sie Gemblere den 16 July. Heden dele morgen artiwerde welet ein Courier, afgefonden door den Heer Keurwerft van
Savin, aan zen Brietanische Majeflege, met tyding dat hy Heer Krurword de Frantien in d'Arrier-Guarde was gevalten, en een Condiderabble avantagie halde gekregen.

And den 19 July. Syn Excellt onfe Couverneur legt met fijn Detachement ong tot Hiven, suifician Beuffel en Vilvoorden. De Graefvan Bergeyk is nu albeen toe Intendans Generael gemaekt, om't Geld
van Bergeyk is nu albeen toe Intendans Generael gemaekt, om't Geld
van Bergeyk is nu albeen toe Intendans Generael gemaekt, om't Geld
van der geraffen. De Spannische Miller ein Erhanden te bestalen.

Gest den 18 July. Op Sondag quarm hier een Frantien Tamboere, om
de pany Frantien, Intellie hier ingebraget, te rantieoneren 1; well gefehreide. Doch is ileut wierd gieberen hier opgelungen, voor wen
de Jarie Fisholen hadden geboden, Na dat de Gerantieoneersden beyten de Foorte waren, delectreerden ist met det haeft van.

Ouffends den 18 July. Ons Connoy in Spannen behended ur;
Schepen, wate onder de Margel van Gent, de St. Javob, 't Nury van
Ouffensk, en de Gulfanaga, is gifteren in Codiex zegi gegaen.

Brafin July 18 July. Sondag prokken 2 Brandenburgte Regimenten
te wate na 't Chaupement van den Starquis de Gaffanaga, du in meer
cen wat van dele Stad leyt, en Macendag 't Regiment Dregorden van

Plain input + Otsu blurred input Pyr her Leger au Gembierer den 16 Ivly. Heden dele morgen steinwerde wedet een Courier, afgelonden door den Heer Keurvorft van Savin, aan zen Brietamische Majesleyt, met tyding dat hy Neer Keurvorst de Franslen in d'Arrier-Guarde was gevallan, en een Contactrabile avantagte hadde gehregen.

Anfil den 19 Ivly. Syn Excelli onse Gouverneur lege met sijn Desactement ong tot Haren, utsischen Brussleit er Vilvoorden. De Gesef van Bergeyk is nu alleen tot Intendans Generael gemackt, om't Geld te onstaingen, en de Spasensiche Militie eile Manedien te beralen.

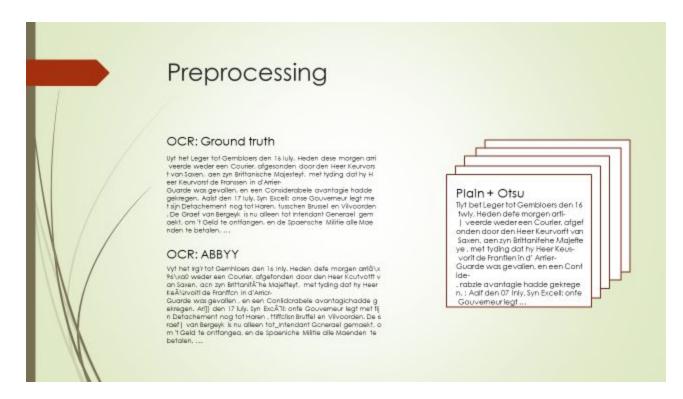
Gesef den 18 Ivly. Op Sondag quam hier een Fransch Tamboer, om de pany Franslen, beelt het ungeforgt, te ransseneren; 't welt geschunden. Doch la Fleur wired gestern hier opgehangen, voor wirst fy al 170 Pilotoe hadden geboden, Na dat de Gestamstungener den buyten de Poors waren, descreereden 'et met det harst tiee.

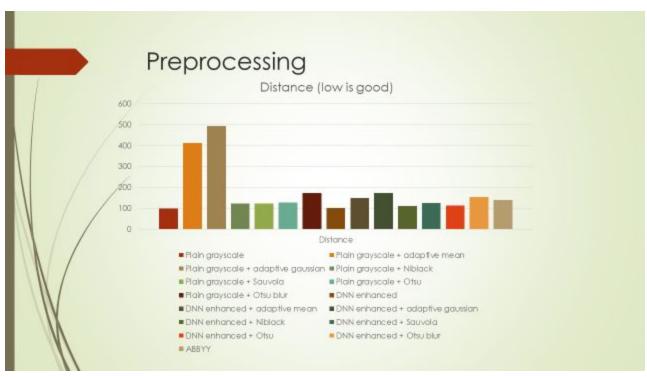
Oustenvik, en de Gallanega, is gifteren na Cadest zeyl eggaen.

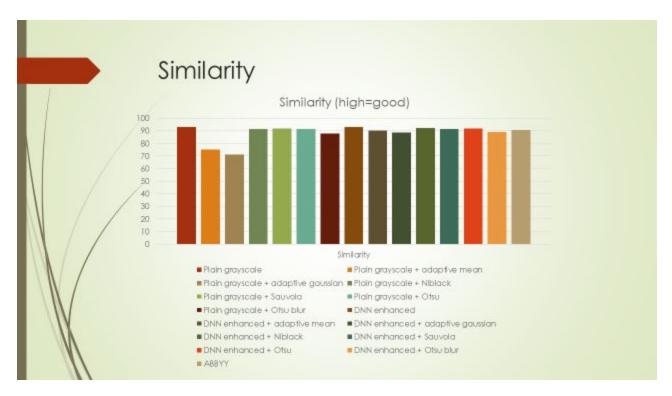
Brassleid was 18 Ivl. On Gomoy en Spanien beloon, 't Huys w...
Oustenvik, en de Gallanega, is gifteren na Cadest zeyl eggaen.

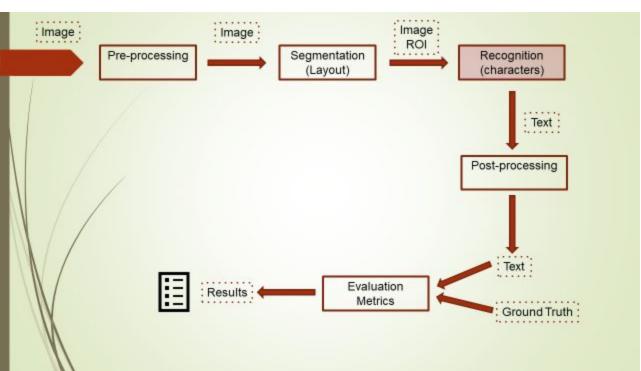
Brassleid was 18 Ivl. Sondag trokken in Branscholper (et Regementer en vott na 't Campement van den Marquis de Gastanega, dat no met een uus van dese Stad leyt, en Macndag 't Regimen Dregonders van

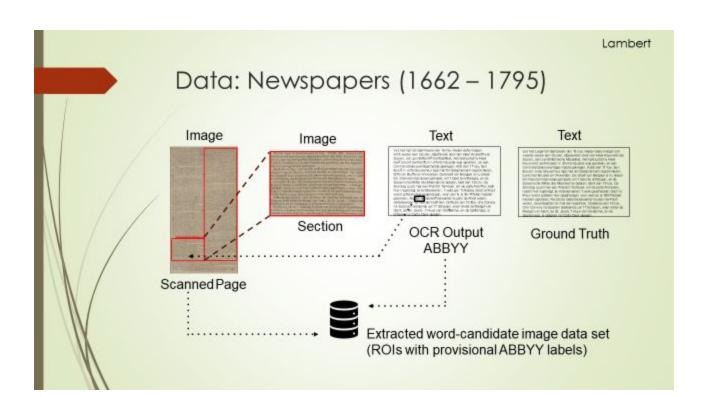
Enhanced by DeepOtsu + Otsu blurred input

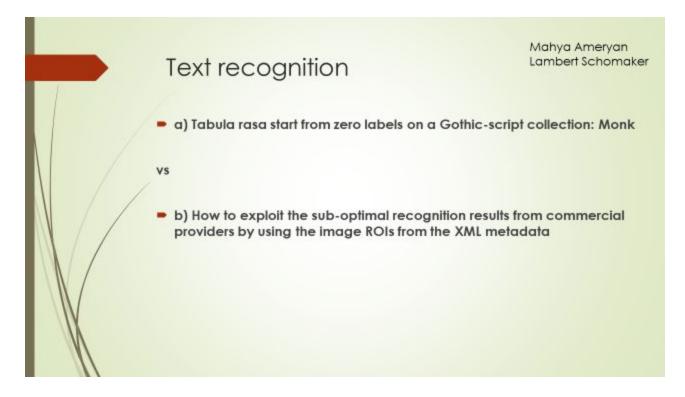










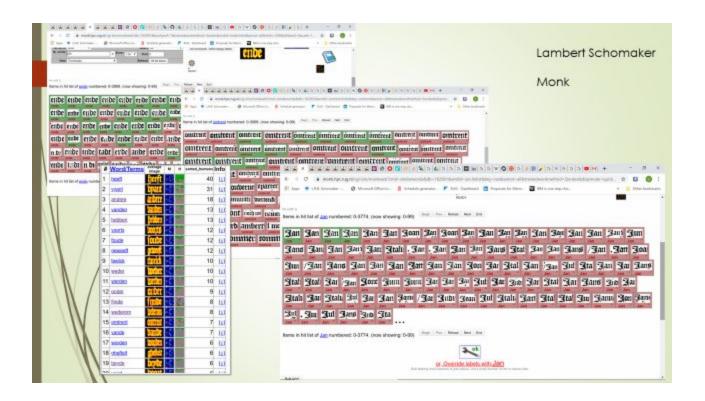


Lambert Schomaker

Text recognition/a - Monk demo

- Tabula rasa start from zero labels, on a Gothic-script collection: Monk
- 436 scanned pages (Geleghentheyt's Hertogen-Bosch)
- XML of commercial provider contains a lot of exotic unicode
- Normalized (flattened) to ASCII to facilitate performance evaluation
- Two labelers, not more than 2 hours of work
- Harvested 319 word classes, 1760 human-labeled images
- 5.5M candidate images processed, learning is going on
- Method: Alternation between a recognition & retrieval engine: 'Fahrkunst' principle "HPC + Human in the loop"

van Oosten & Schomaker (2014). Separability versus prototypicality in handwritten word-image retrieval, Pattern Recognition, 47(3), pp. 1031-1038



Mahya Ameryan Lambert Schomaker

Text recognition/b: exploiting results of 'commercial' recognition attempts

Methods of labeling:

- Human labeling: high quality, but expensive.
- Machine labeling: Maybe low quality, but massive data

How to exploit the sub-optimal recognition results from commercial providers by using the image ROIs from the XML metadata?

- Using the 'Meertens' data set:
 scans + sub-optimal recognition results from the Alto XML format
- Collect image ROIs with word candidates+ their provisional labels

Data & result of Word recognition

Mahya Ameryan Lambert Schomaker

	#Images	#Classes case sensitive	#Classes case Insensitive
Train	10,037	701	654
Validation	3,346	665	618
Test	3,346	665	618
Total Dataset	16,728	736	689

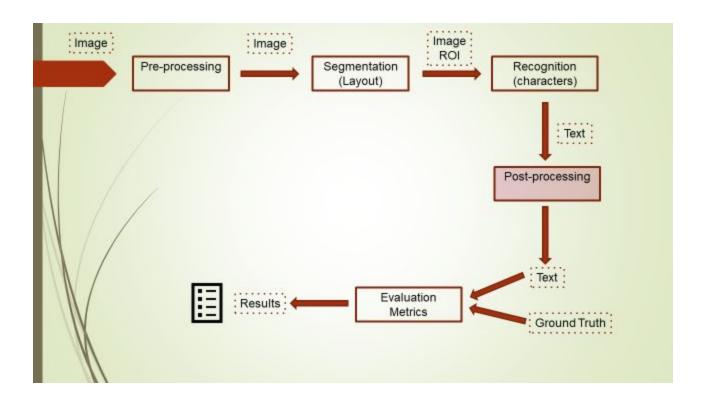
- Alphabet size is 66; it includes [A-Z] [a-z], digits, and ^? = -.@
- A CNN/LSTM model was trained
- 88% Word-recognition accuracy

(while 100% can never be reached!)

Mahya Ameryan & Lambert Schomaker (2019), A limited-size ensemble of homogeneous CNN/LSTMs for high-performance word classification, arXiv:1912.03223

Text recognition/conclusion

- In current practice, the ROIs (images) of recognized words are never used
 - Our results show that this information is highly valuable, even with errors.
- Next steps:
 - Cleaning up the data: Filter out strange OCR results (words start with letters) and assume that a word string that happens >n times is not a coincidence: it may be a lexical word (n=5 or 10).
 - Applying the method on a more realistic 4894 (5K) classes
- Raw ROI-image extraction from .JPG/.XML is ongoing: 22k classes, 78k instances on 24/1/2020, 4:00AM, to be recognized later



Konstantin Todorov Koen Dercksen

Linguistic post processing

- Problems in OCR can be solved on the input side, but also on the output side
- Goal: to improve garbled digital text output from OCR to something that is closer to the target language

Konstantin Todorov

Deep learning approach

- BERT state-of-the-art in current NLP tasks
- Transfer learning:
 - Use pre-trained BERT
 - Fine tune so it learns Old Dutch
- Machine translation model
 - Assume uncorrected and corrected as two different languages
 - Train a neural network to transform uncorrected to corrected text

Translation goal:

Uncorrected

Uyt der MofÃ"ou heeftenen van den 14

Juny, dat den gednrioch brandt in die

Stadt continueerde, apparent

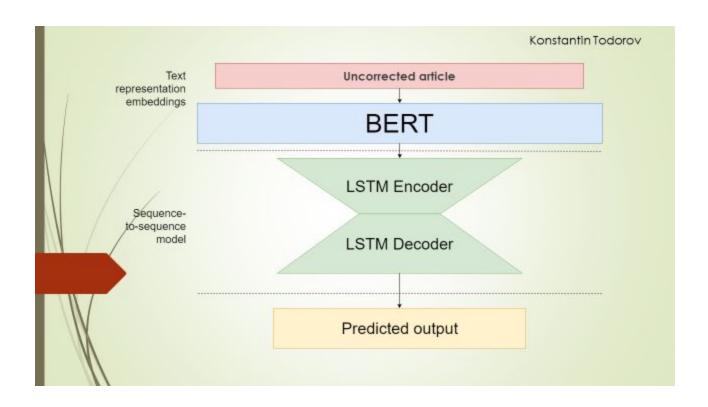
doörmoortbrandersaengflftieht, op den
n Juny branden alleen 5000 Hiiyfen, vecle

Ammunitie-, ende (f 44 menfchcn.opden
11 Juny £000 Huyfcn met vyf Czuarfc

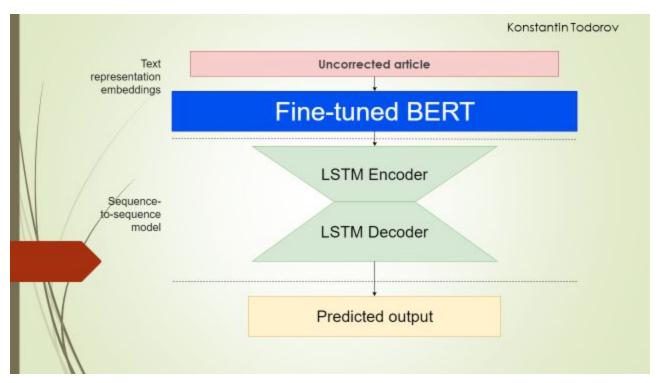
Cancelyon endeeen groot getal

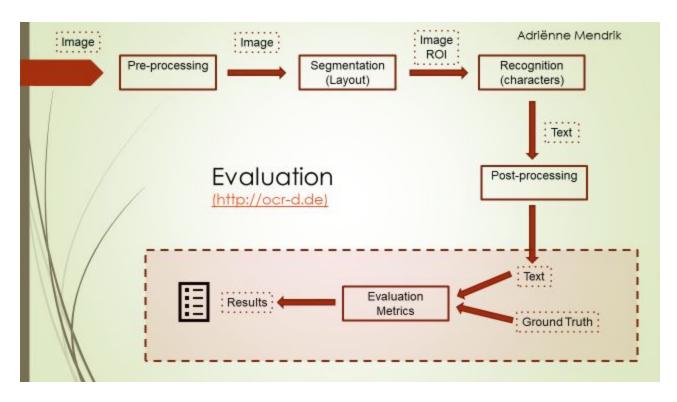
Corrected

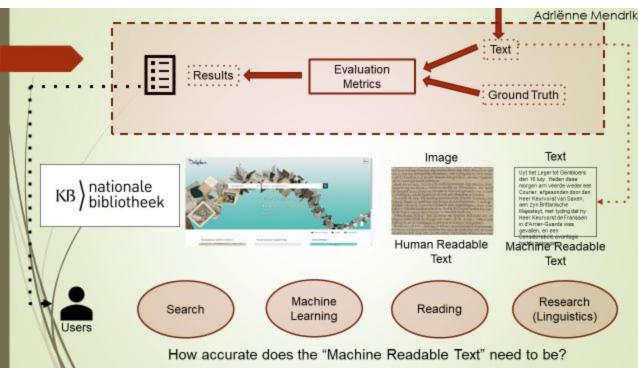
Uyt der Moscou heeft men van den 24
Juny, dat den gedurigen brandt in die
Stadt continueerde, apparent door
moortbranders aengesticht, op den 12
Juny branden alleen 5000 Huysen,
veele Ammunitie, ende 644 menschen,
op den 22 Juny 6000 Huysen met vyf
Czaarse Cancelyen ende een groot
getal menschen

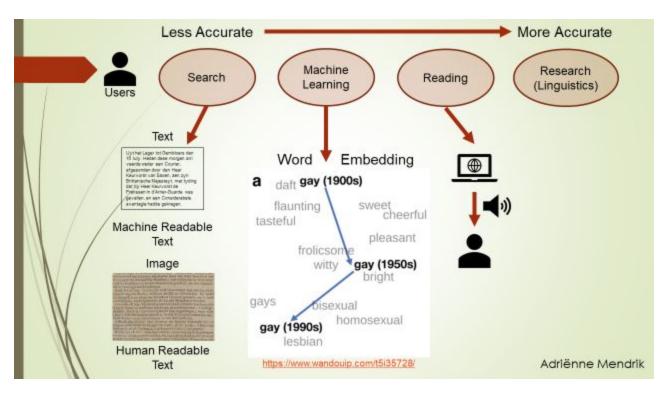


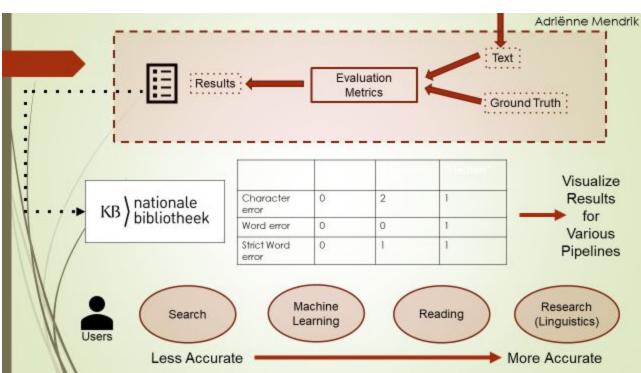


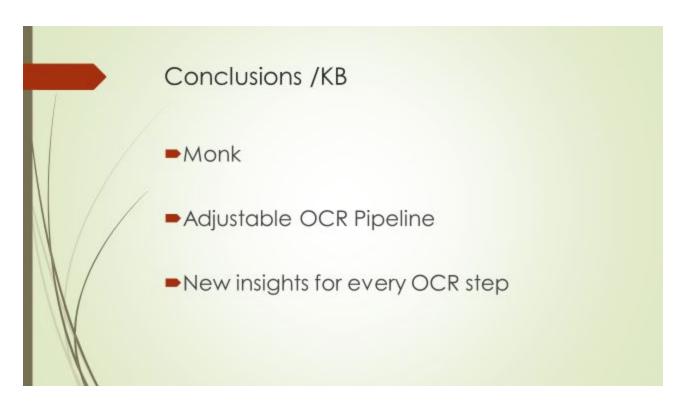














The near future...?

en waeren gedzuckt/ tot Bzussel by Jan Peiker vande Stadt/ en de France tot Loven ende Coz. Coenestentum/ende bevonden daer ken/ int Latingspestelt van Cornelio Iansen sen Tytel: Alexi Pharmacum civibus Sylvadversus ministrorum suorum fascinum, end rentie binnen der Stadt/ ende vooz de Burg plat afgeslage wozde: niet te min dat dese im dispupt / ten platte lande te treden/ op een Frontieren/ die vyp is vooz bepde de part boozs. Pzedicanten geantwoozt ende int geh

en waeren gedruckt/ tot Bruffel by Ian P ker vande Stadt/ ende France tot Loven b ende Cor. Coenestenium/ende bevonden dae ken/ int Latijn gheftelt van Cornelio Ian fen Tytel : Alexi Pharmacum civibus Sylv adverfus miniftrorum fuorum fafcinum, en rentie binnen der Stadt/ ende voor de Bu plat afgeflagë wordë: niet te min dat de in difpuyt / ten plattë lande te treden/ Frontieren/ die vry is voor beyde de par voorfz. Predicanten geantwoort ende int

KB national library of the netherlands

KNAW Humanities Cluster

Delpher

Thanks to the team!

