



Open Source Handwritten Text Recognition on Medieval Manuscripts using Mixed Models and Document-Specific Finetuning

Christian Reul¹, Stefan Tomasek¹, Florian Langhanki¹, Uwe Springmann²

> ¹University of Würzburg ²LMU Munich

> > 25.05.2022

Introduction – Model Types and Where to Use Them

- Different approaches and models for different use cases
- Model types
 - Mixed models (MM): trained on a variety of books / manuscripts / printing types / hands / ...
 - Document-specific models: trained specifically for the recognition of a single book / ...
- More specialized models yield better results but their creation requires more effort
- Example use-cases
 - Mass text recognition, e.g. in libraries or archives
 - Manual correction not feasible
 - Apply MM and hope for the best
 - Preparation of a critical edition
 - 100% error-free text required \rightarrow high amount of manual correction inevitable
 - Recognition \rightarrow correction \rightarrow training \rightarrow recognition \rightarrow ... until everything is transcribed
 - \rightarrow Gray area in-between



Use-Case – Overview

Digital edition project "Die Kindheit Jesu" (Childhood of Jesus Christ) at the University of Würzburg

- Material: Medieval manuscripts written ...
 - between 1250 and 1350 in Middle High German
 - using two very common German medieval writing styles: Gothic and Bastarda
- Goal: Comparing text variants using a configurable filter system based on different distance measures
- Some distance measures, e.g. Word Mover's Distances, can and have to be trained
- Training data needed: Non-normalized text in ...
 - (very) good but not perfect quality
 - large quantities → efficiency is key → combining mixed models and document-specific finetuning

Julius-Maximilians-UNIVERSITÄT WÜRZBURG

Use-Case – Task and Workflow

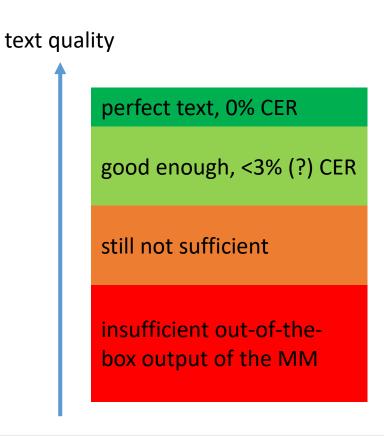
Task: Get from the out-of-the-box output of the mixed model to a good enough CER in a very efficient manner

Workflow:

Julius-Maximilians-

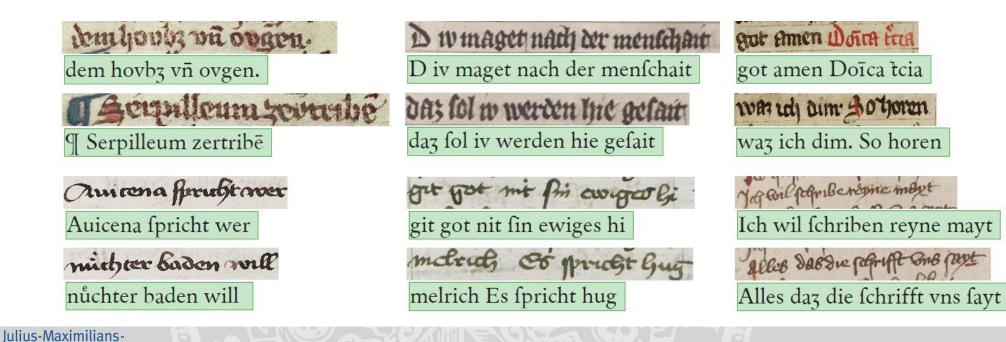
- Train suitable mixed model(s)
- For each manuscript:
 - 1. Apply the most suitable MM
 - 2. Manually correct some pages
 - 3. Train a document-specific model using the newly gained GT
 - 4. Apply the new model
 - 5. Repeat steps 2-4, if needed (Iterative Training Approach)
 - 6. Apply the document-specific model to the remaining pages





Data – Mixed Model Training

- Suitable manuscripts from ca. 1200 to 1500: Kindheit Jesu, Marienleben, Parzival, ...
- Self-created or compiled from freely available sources
- Final corpus: 35 works (at least 30 different hands), close to 300 pages, ca. 12.5k lines



IVFRSITÄT Open Source Handwritten Text Recognition on Medieval Manuscripts Christian Reul, Stefan Tomasek, Florian Langhanki, Uwe Springmann

WUR7BURG

Data – Preparation

- Fully open source workflow!
- Preprocessing via OCR-D (<u>https://github.com/OCR-D</u>) processors
- OCR4all and LAREX
 - Segmentation and ground truth production
 - https://github.com/OCR4all
- Calamari
 - Recognition, training, evaluation
 - https://github.com/Calamari-OCR







Wá a colonnet ain wib Wå er erkennet ain wib Wå er erkennet ain wib

Die nit verfagen kan Irn lib Die nit verfagen kan Irn lib

Da ÿlet ez wil balde hin

Vnd wirbet vaft daz ift fin fin



Data – Evaluation

- Four manuscripts (2x Gothic, 2x Bastarda) written by previously unseen hands
- Kindheit Jesu, Marienleben, Der Welsche Gast (x2)
- Each manuscript further divided into train (32 pages) and eval (8-18 pages) subsets

Daz ih des niht vol bringen chan.

mir chom zehelfe tar an.

man tún fol zú allen zýten ton man tůn fol zů allen zÿten. vñ

wå von man nit treg fol fin vnd

- oaz buch bræhte her zewege daz fie 12 alle daz buch bræhte her zewege. daz fie iz alle mufen lesen die gotef kint wellent we mufen lesen. die gotef kint wellent we

Daz er begang mit guter dat

waz er gûtes gelefen hat

Julius-Maximilians-UNIVERSITÄT WÜRZBURG

thar verbuzzer wol ten wil. 1º drevithe levhter ober al. v ful with it' gettowen - Jus wan with oue fonde bowen, as fi vns ellende Poluch beim sende. nde bas finverfe v cem fronen paraderfe. to a fire eva our servel on tem mere chan fi ten duelarre wol geterten o zoufen arebeuten. over hund porten. in it den gotes worten. 1 mart Div marbeit gefagt. t if muter vn mage. t hat ter engel niht bewogen. or that was forme gelogen. b m- vil edelen protte-To the me mannes geluit Tervan noth ver funden val. 149 mirer and mal

Julius-Maximilians-

UNIVERSITÄT

WÜRZBURG

men lan merbens der evangelitte nam m to lange ir die untze ez vil reine gar er to re daz weder mot noh more uni worre enmach geuruben des lar öch wh genn gen sin fant terommi lere die mugen w gehelten an ör tele.

1 den telben enen got enenge unfa clumchely wurth wan eme erch ant muid ca fremide was eran Dertiva do was der heidenschet Te vil, die des meneles ibil ander, alagona i begreugen des och fie faden geviengen miom libe vne ander fele dobelnett die gotes c. div mahelifke diet. als monogles genet vud ir viter babraham Sin waren. Ge achoriam staac vno meob banals, der innze inmele fahrem lener non Sevende gerine da got zetiner getine, selbe wider minah winders un öh mere refeats de er ben engel gevie:unt in des more frie he mule er in gefegenor die fer in por da in Spengel or ular him and the gerulne: zeemem mehunde. hinener on

Open Source Handwritten Text Recognition on Medieval Manuscripts

Christian Reul, Stefan Tomasek, Florian Langhanki, Uwe Springmann

30 Den man more wunden Die wee herren frane haz na wart er inchuizen funten. vnt uhen er geviel in baz. ver herrzen lærr & fucher har 9200. Ganz vi ane malenhard in finar nunder ein treil. Weder for noch hant. erra Mut tem mil ib erzwgen diz. ald rener as emphant. dur einem fei mit vribten biz. would er of fprane. Der ungemaches uft gewon. 2205 fil hab unmer danc. Senne tenen & at von. Arowe daz ov mit haft ernert. Nie ceben arbant. vil tem grunme tob erwert. 2215 fo vil fo in einem tovin erlant man ih mertich mus erflagen. -Daz Difem ein fenfte mere. Dem rive on herrzen fuere h fprach ov haft genate lagen. 2210. Gove ter dich erlofer hat. Trownen wnt fenen gtt. valer bund infeine hante fat dar doubte ienen en linuzit-Er uft tein anzt gewelen. 22.50 to en me nibe lances seldal wand er gemach for ungemab du bit benamen wo un genefen-1 wount dw tor of getan-Im coult wol not we. vi iene mit vrive in vlan. 2215. fus prif ih feme write me Dieman e-Automunde wa tracip ber zemazze chumb Lanot. ungeringer soh mbr belæn - 2255. Danne den er gar vermanter Az felsen andifem gefeile he-Ob er genelen wore of wie . no officien ez de looffionenie. wit an dem wurte do fie-2220, Awiter einem worte dur an-Ir samer ab wunten. wie fi die falben genum. vor frivæn fine chunten. ezhal och enwellen we gehmen. v baten die fchahare. allo mouride fi e maren. ob fi marele man. wol bunterfunt als with Si gebe fer m vit nam u babe wunden fi ur gelukes to. soft founds as wenter web abe. To teneden fil die fehalman. Wir gewinner des tehemen mit S wirt belaup iene chiten bin an acharin man fo groz quit 208 Emer her ter anter hin. Zeder wertte ie a wunne. fi wrfen vmbe cen gewin tes ih vin al men divine vlerchluchen als e. 2230 mmer gericht ware. mag wahte umm me Sarvube ib u colume. belle biz fi funden. Warz got ihn velouf while. eero Das wit ses fi arden den dumten th wil frem wourfr getdilt. mat zenvertes eren todate. Damit fishen fein Leben. win in gewilligen mochte 2205 Jurh in & mur fihur gegeben.

.

Bastarda

Euclid win concept an erite fan Vit vol an autor false an Vet ger von ger gite fine Avnes hielier itt fie bernine

Or for gebort and grelefen 1 this worthich biles man Bl Fu allen zönen finerhen wol Oder nin oder gedouch Non Sem wetter place mit anarchen Map aft imper hires ontugerst Dearfait simpt sit wol by incier Warm man not Bu tim Rat Wan hab den fin verd orth Son wit Dag man cunveler forestic wol (were thibble wil fin and profinge De growinet prove geninge Watassie ander devien Drington per may tax an vilevel gelingen Were her inget like might hope Manne ex wit en solte Da er morter dag er felte

Wet an onquelle for suger sweder Der har fin also par putionder Wer alexavit mit even like Der Plunde even hin unfer freeben Wan bit plan die mougent When man fac an frate molec wget Man bet abjet nit mere ift Bo feet Bas Bind And de feel Inden Ramqueten fin In fee Bin welaft wiet michelt mer Dem faller wit Day failin hip Mann or mine fat wiffer day Tem fight of most offen not per winder of nach winder to Man er mit nu voinden Gar Be willers book wit Raken war Alfan dem alben manne gafdpacht be tan fich entifaben niett Dex consigne work der withight Des ce oftag in fince in gent Da ve port ing in while Das min fine unget wool leftate Bun fichtige Bast von an grute dingo Mone under Sacan sit millingen Jefuens den hinden Dife Cece Offic in San inft wollet more Das migen fie Davnach permuen Ore fallent fiel plane 30 mapen Norm war fiel plane 20 mm polision Norm Lugo por ond palletinit Ond monopolar confectiont An driven Diggen ma gaber pol Setume aver is wil pofferent wool Time ton man wir forethe sonere The many goline welfer on wel

havit day dem ubelin manne The making . And day man first beffer wind dem iller manuscrist volge /31 1) Is preself ref wine ref des . Win. Dute Gab wax openome Soz alle Sie Die mit Godffort BRimeet fin son matter wie man dort der lighten Luke menie wave Ond forces and er alle Gort fortig fi dax fine genn nit carl condetaning fin als dar der andere Gerfiftet Dan er fol. Da Das mir der geon condectioning Bellen fin die voo got 34 Geon Gaurgebed (2nd by wie wo den frientjon Gilde mieme. Sie des mit woolhen roolgen Bud onle orf des forthen Bold: an navcone Agrocher Mid to andern Mider alter E. Din wie das gefigigt Day ma Durch funde drean einen Sofen Gody Baur Rudwie wir hin mit rolaren wille ober with boloch if One make org de G.D. man fic Edwande more peffection for Ont. Grange Sam Day cont yot ame maiplex Bat opposid don Atgelie maine alle sit. Tog maine de Babit This wie day doe hit der hu micge : ache Oud forer wie uset Das Di der ballt fine bedier on fine Brieffe pende Indes Gailige grabes Dient On dete Das Querts opin And owin das que user Beat Min Page orgroo dem quiten Frierigte dor Da To ang Du wie man sin put Ding high rouriere may abrain Boles mit policit (200 wie lichte

merate may Sarma quer hit Mind Great ain went too den Corseen. bus in Darmink of for and Day since Sem and mit entratest norm Dot expern bit. The wie fin wnour ffe fo Das ma wider wite Das ma uter mere montes dienft refit vac. warm ma tites doch congress is a mans instotes name aler varn Na gaine In maning works barume fie co un follen . This anterviete och ben Sie lave wordere fipreceffen . 2. Drie mans with Sie fingten to den subfichen land en Day fie Day felle tim Day fie uber mer valen And poge sel was fining-feiderief da werten per of fin got werteger sei Da prieste ief wie up up miner matery Fumon Em ani wenig ven wit wider Dach An waraunt die Bort fait als fraik: Are wie and to doc Gott fact on finif enden veallen mift. Wid gaine Dan mit welden lifen ma Die unternet Abierthen al Vin Dit name die Gorth Fact - Jand Die engo vin den mide And tax succept and die maindert Ic mathe self ani . wenig rook we Whit page wie min fider Glagett da fie zu" Answücte von Role Dann min Guil an - n. W no pocutic das an fice Som aren von dem Lowen zolaren BENEW BE ainen annow ame Geowen Infine mult beggen weder minute work mere 201 mar zwifth

I Som fit out day best effor me Eman bringe Die ersterentigt wanne en mit legen pot wort er fich behinten wort Berdeurhallen (In Som munt Ar behare fieldouch mit & Der fime Die Wile auf m dom munde har refit Anemon lieve fich 31 Den defetten mit dem berhel to fie winthen weillent E fie in sperime tool Som munden Det vom har fie dar zu gebunden Dog Bimpt Bouchelien Lites mehr Ein man vor den epefetten fin soie affen das aft die here mins Obe ine So right gelatte that So im fother er effen pl mon plont of y atte fift mir So home Sie engegen aft Buy Der gefette Bi Der verhren hames Man pot and Day open rounday Day man mit effe mit Beday franky man st ouch gaben meht Day man mit dan gemaffe nigt (In die philled gerffe mit Der hanne

Julius-Maximilians-UNIVERSITÄT WÜRZBURG

Experiments – General Methodology

- Using Calamari version 2.1 and mainly its default parameters / best practices
 - Five models per voting ensemble
 - Five augmentations per original line
 - Early stopping after five epochs without improvement
 - ...
- Evaluation
 - Apply models to evaluation data
 - Standardize the output analogously to the training data
 - Calculate the CER using the Calamari evaluation script



Experiments – Training Pipeline and Preliminary Experiments

- Training pipeline to successively refine the models
 - 0: Starting point: LSH-4 (strong MM for material printed between 1450 and 1900)
 - 1: Utilize the entire training set (Gothic + Bastarda)
 - 2: Train two writing-style-specific models by exclusively using Gothic/Bastarda data
- Summary of preliminary experiments
 - Application of in-domain MMs (Gothic \rightarrow Gothic and Bastarda \rightarrow Bastarda) yielded best results
 - Out-of-the-box application: very clearly
 - Starting point for finetuning: in-domain > combined > out-of-domain >>> starting from scratch
 → used in-domain MMs for all experiments
 - Calamari predefined network structure *deep3* performed best → used for all experiments (cnn=40:3x3,pool=2x2,cnn=60:3x3,pool=2x2,cnn=120:3x3,lstm=200,lstm=200,lstm=200,dropout=0.5)



Experiments – Simulating an Iterative Training Approach

- Iteratively doubling the amount of training data
 - 2, 4, 8, 16, 32 pages (4 subsumes 2, ...)
 - Starting point: the respective most suitable MM
 - ootb: out-of-the-box result of the MM

Julius-Maximilians-

- Evaluating on a fixed evaluation set and calculating the CER; averaged over all manuscripts
- Decent ootb CER considering the material and MMs
- Training on just two pages cuts the CER in half
- CER of 2-3% already completely sufficient for our use-case, many downstream tasks, and efficient further transcription

# pages	CER in %	Impr. in %
ootb	6.22	-
2	3.27	48
4	2.58	21
8	2.17	16
16	1.94	11
32	1.65	15

Example Results After Finetuning on Just Two Pages

uon fo heiliclichem fite. gab im got finen fe uon fo heiliclichem fite. gab im got finen fe uon fo heiliclichem fite. gab im got finen fe

gen. himeltov und regen. daz ez michel gegen. himeltov und regen. daz ez michel gegen. himeltov und regen. daz ez michel ge-

nube gewan. fwaz er buwen began. an fwr nuht gewan. fwaz er buwen began. an fwiv nuht gewan. fwaz er buwen began. an fwiv

fich der herre uerlie. mæheil ez fur fih giø. fich der herre uerlie. mit heil ez fur fih gie. fich der herre uerlie. mit heil ez fur fih gie

 $\bar{\mathbf{v}}$ ergab un fælicliche. er was alfo vibe richo. $\bar{\mathbf{v}}$ ergab im fælicliche. er was alfo vihe riche. $\bar{\mathbf{v}}$ ergab im fælicliche. er was alfo vihe riche. Er erwirbet amit rům minur daz Er erwirbet mit rům nur daz

Daz ain ieglich man daz gicht Daz ain ieglich man daz gicht

Daz fie es tet durch ain boswicht Daz fie es tet durch ain boswicht

Des ift fie wol ain bofes wib Des ift fie wol ain bofes wib

Daz fie geleftert hab Imm lib Daz fie geleftert hab Imm lib Daz fie geleftert hab Irn lib diff view shows deletions and insertions necessary to turn the prediction into the GT

line image

ground truth

diff view

Example: model predicted "m", correct would have been "rn"

Experiments – Effect of Mixed Models

- Using MMs as starting points compared to training from scratch
- Unexpectedly vast improvements ...
 - when using small amounts of GT (85% for 2 pages)
 - but also for large training sets (39% for 32 pages)
- Suitable MM not always available or not strong enough but ...
- ... even starting from an out-of-domain MM yields large improvements (>50%) over starting from scratch

# pages	without MM CER in %	with MM CER in %	Impr. in %
2	21.12	3.27	85
4	10.74	2.58	76
8	5.82	2.17	63
16	3.78	1.94	49
32	2.72	1.65	39



Conclusion

- Using a fully open source workflow we ...
 - created a training corpus covering two widely used German medieval writing styles
 - trained two performant and <u>freely available</u> MMs
- Document-specific finetuning quickly led to very low error rates
 - 2-4 pages of training material already enough to reach our goal of <3% CER in most cases
 - Use of suitable MM as a starting point improved effectiveness and efficiency considerably ... but even a less suited MM is better than nothing
- Future Work
 - Improving our MM as a side-effect of our data collection
 - Opening up new domains by applying the same procedure
 - Automatic quality estimation using Calamari's intrinsic confidence values

Julius-Maximilians-UNIVERSITÄT WÜRZBURG