More Data - More Problems? Die vielfältigen Herausforderungen für Automatisierte Texterkennung auf Dokumenten (nicht nur) des Herder-Instituts

Ole Meiners Arbeitsbereich Forschungsdatenmanagement Abteilung Digitale Geschichte und Informationssysteme Herder-Institut für historische Ostmitteleuropaforschung

DIE VIELFÄLTIGEN HERAUSFORDERUNGEN FÜR AUTOMATISIERTE TEXTERKENNUNG AUF DOKUMENTEN **DES HERDER-INSTITUTS**

- Ziel: Automatisierte Erstellung von Volltexten zu Beständen des Herder-Instituts
- ... als Grundlage für weitere NLP- und DH-Verfahren (NER, Topic Modelling etc.)
- Bestände sind...
 - sehr heterogen (zeitlich, Medientypen, Inhalte, ...)
 - vergleichsweise kleinteilig -> hoher Arbeitsaufwand GT-Erstellung etc. für ,maßgeschneiderte' Modelle in Relation zum Output
 - nur in Teilen und nicht systematisch digitalisiert

26 Smil van 4 Systember 1863.

Links Doles: - May like is in later 24 Hours von Grofo is is fell in harfriell gelan, and son, Vago jo regames Eligar Erflas Sarah, nesas in dar hinkorns, mas jo nafranas ift. - Of fefor Januar nors Wajafor, rayorm is is Rinter la frift falls, it's alla his not bledwig befor mit reason. Coffers Jahan fis faiflest suffield not about Jeggiert, for Winterwolasten beganna, N. J. mil viballones jagisfell de If fall glirkling Maricas Tofiles and sho Wagan bringen i Joffe, rep iso Bacron Year Morales & Viletter nogerhroufen ratoman dean't falls of the Gracier griff nes nyarla mis star so fregulary var Riflar in the formers Eister the Chychianalin man fis wher my allas markovanto Kens you asfalfan si sal tifalgabases in stifan Valor anjenhuras notyvarjar. - Talas Walls fif favor favors, rafe itis, venan Holefanfar quallet renoven enerna, Nen fin 3nd New allysbrancher workfield las y wester

sid alle chne Fensterglas, chne Turen. Durch Genchesseinschlüge und desclierte Ofen und Winde sind die wertvollen
Bücher mit Adselbsteinen, Schutt und Schnes bedeett. Tusen
dem eind diese Rüume als Abert benutzt weden micht Bicher
aus frühreren Jahrhunderten das mitige Burlen und die eine Schlegen schlenen. Se erweist eich dass beim Sinner und bester Ordnungen
die unserbiert kostbere Siblichten noch in bester Ordnungen
Begalen und in Schrönken geordnet gewesen let. Die Spanier
haben aber alles was Hols war verheigt und die Bücher Teken
haben aber alles was Hols war verheigt und die Bücher Teken ausende von Büchern eind aus dem 18ten, 17ten, vielleicht Sten Jahrhundert.Gewiss ist es möglich, dass die Sowjets selbst schon einige Bestände rechtzeitig in Sicherheit einiges herausgesucht hat ... Zauch sollen, wie wir hören, der spanische Divisionsprediger, Offiziere und gebidete Soldater spanische Dry. monspranger, wilniere und geriaete Sciaere der "Blauen | Delen"dich de schom manches kotbers, gute mit genomen haben, sie fünden neben ruseischen, deutschen fran-zögischen und englischen Nerken viele lateinische und segar spanische. Wir stellen aber auch japanische chinestache, turkische, arabische und wer weise was für Schriften fest. Sin Raue enthält nur elawonische, vornehalich geistliche Litera-

Braussen im Schnes finds ich eine Menge von prachtvollen russischen illustrierten Werken, aber sie sind alle nass ge-wesen und jetzt zu Eisklötzen zusamangefroren...unrettbar

Dr. Wunder spricht mit einigen Spaniers, die hie im quartie etehen,bittet,emm wöchte die Büchervie: den vernaceln, und vor weiters Verunreinzung echtbans. Unser Sett ist benes-sen, wir können selbet nicht echt tung, wir könnennur berich-ten und veranlassen. Jud das wird geschähen.

(Auf unders Beranlassum him ist sehr bald darauf von einem Hehrmachte-Somieskomende mit LEM's die ganne erstischtelliche stelliche meh Bien abschaft worden und dort der Obhrtyfflesund forwaltung des obersten orthodoxon Gedetlichen au treuen Himles übergeben verden; ich hoffe, dass die syfter von den Sowjets wieder nach Howgored sarückgebracht worden ist!)

Se ist schon nahesu gans dunkel, wie wir des Areal verlassen. Vanittelber jenestte des grossen Tores muss der Fahrer den Nages nochmal enhalten und die Notorkaube aufhöben. Ich espa-niere herne und ziehe swei starrgefronen Sachen aus des Schnee, aus des ich des swert hernunkelen mussel auf Schnee, aus des ich des swert hernunkelen mit bei füllt air els goldig schlassert auf se ist weldfranen versierte, kroutförmigs bromteilardecke. Hicht weit davon lugt eine hellbrauner ledernen Bachdockel aus den

DIE VIELFÄLTIGEN HERAUSFORDERUNGEN FÜR AUTOMATISIERTE TEXTERKENNUNG AUF DOKUMENTEN (NICHT NUR) DES HERDER-INSTITUTS

- Ziel: Automatisierte Erstellung von Volltexten zu Beständen des Herder-Instituts und anderen **Textdigitalisaten**
- ... als Grundlage für weitere NLP- und DH-Verfahren (NER, Topic Modelling etc.)
- Bestände sind...
 - sehr heterogen (zeitlich, Medientypen, Inhalte, ...)
 - vergleichsweise kleinteilig -> hoher Arbeitsaufwand GT-Erstellung etc. für ,maßgeschneiderte' Modelle in Relation zum Output
 - nur in Teilen und nicht systematisch digitalisiert
 - rechtlichen und ethischen Beschränkungen unterworfen

Sufil van + Vyslembes 1863.

Links Colis: - May hip is jo lains 24 Minister von Grofo is is fell in harfriell gelan, and son, Vago jo regames Eligar Erflas Sarah, nesas in the historias, mas to reform ift. - lef fefor Januar nors Wajafor, nayorm is is Rinter la frift falls, is alla his not Hedrig befor mit nernan. Coffers Jalan for fingling suffield not about Jugiert, for Winterwolasten brganna, N. J. mil Vihablones jagisfell de If fall glirkling Maricas Tofiles and sho Wagan bringen i foffer, rep iso riverfor year. Morales & Viletter nogerhroufen ratoman dean't falls of the Gracier griff nes nyarla mis star so fregulary var Riflar in the formers Eister the Chychianalin man for wher they allas markovands horn you asfalfan s. sal Sifalfahars in stifan Valor anjenhuras notyvarjar. - Talas Walls fif favor favors, rafe itis, venan Holefanfar quallet renoven enerna, Nen fin 3nd New Whysbrander workfrifto las y wester

sind alle chas Feneterglas, chas Türen. Turch Geschosseinschläge und demolierte Ofen und Wände sind die wertvollen
Bücher alt Maggalsteinen, Schutt und Schnee bedectt. Tusses
dem eind diese Bümen als Abort benutzt worden, wobel Bücher
auf Früheren Jahrbunderten das Bötige Tepier zu liefern
schlanen. Be erweist alch dass beim Sinau unserer Truppe
die unserhört kostbure Bibliothek noch in bester Ordnung us

Annahmen und in senteninen zerofent Enuspean jetzlie Spanier einiges herausgesucht hat... Much sollen, wie wir hören, der spanische Divisionsprediger, Offiziere und gesidete Soldater Saischen und englischen Werken viele lateinische und sogar panische. Wir stellen aber auch japanische, chinesische, turkische, arabische und wer weise was für Schriften fiest. Sir Raue enthält nur elawonische, vorneholich geistliche Litera-

Braussen im Schnes finds ich eine Henge von pracht vollen russischen illustrieren Werken, aber sie sind alle nass ge-wegen und jetzt zu Eisklötzen zusamsengefroren...unrettbar

Dr. Wunder spricht mit einigen Spaniern, die hier in Quartie stehen, bittet, mam wöchte die Bücherder doch vernageln, und vor weiterer Verunreinigung schüten. Unser Zeit ist beseben, wir können selbet nicht mehr tuen, wir können ur berichten und veranlassen. Und das wird geschehen.

(Auf unears Veranlassung him ist sehr bald darauf von einem Hehrmachte-Sonderkommade mit 1878 die ganse erwischelflicht mit in sehrendt worden und dort der Obbut; Ticht mit in speechat worden und dort der Obbut; Ticht mit und vermitung des obersten orthodoxon Geistlichen zu treuen Elinien übergeben verden. Ich hoffe, dass ein spitter von den Sorjets wieder nach Howgered zurückgebracht worden ist!)

on 124 School mandau salm uniter, was war des Arent vermande venittelbar jonestie des grossen Tores unes der Fahrer den Reges nochmal ambatten und die Notorhaube aufheben. Ich spa-niere herum und ziehe zwei starrgefrorane Sachen aus des miere narus und missie swei beniege intene cachen alle des Schnee, aus des ich sie muere heruubackern mussidas dies Hillt dir als goldig schissenrd auf; se ist die klaine, mit Woldfransen vermierte, krousförmige mychieklardeckei, sicht woldfransen vermierte, krousförmige mychieklardeckei, sicht wolt davon lugt dies hellbrauser ledernen Buchdockel aus des

- Auswahl von zwei exemplarischen Dokumenten für HTR/OCR:
 - Briefwechsel Julie und Eduard von Oettingen, DSHI 190 Livland 33, 11c, ~3000 Seiten (Handschrift 2. Hälfte 19. Jhd., siehe Abb. 1)
 - Georg v. Krusenstjern: Einsatz. Tagebuchblätter, Briefe und Notizen aus dem zweiten Weltkriege 1941-1945, DSHI 190 Krusenstjern 2002, ~400 Seiten (Typoskript Mitte 20. Jhd., siehe Abb. 2).
- Erstellung von Evaluationskorpora (50seitige Samples, händisch transkribiert)
- Erprobung von off-the-shelf HTR/OCR-Modellen u.a. mit Transkribus, eScriptorium, ocr4all, recognAlze, nopaque, PERO und Abbyy FineReader
- (Nach-)Training/Finetuning von Modellen zur Layouterkennung für Handschrift in Transkribus
- (Nach-)Training von OCR-Modellen für Typoskript über eScriptorium

26 Smil van 4 Systember 1863. Links Bilis: - May light is is likes 24 Hindres von Grofo i is fell is Parfriell gelen, and sous Vago jo regamin Elligar Erifles Thanks, was as in dar hinkorns, mas jo nafranas ift. - Of fefor lankan nors hajafor, nasform as son Rinter la frift falls, it's alla his not bledwig befor mit nernan. Coffers Jalan for fingling suffield not about Jeggiert, for Winterwolasten brganna, N. J. mil Vihablones jagisfell de If fall glirkling Maricas Tofiles and sho Wagan loingen i Joffer, rep is is is acres years. Maintes & Vilettes nogerhoofen sakouan Chart falls of the Graciary will now nyarla mis star so fregulary var Riflar in hib frams Eister the Olychrenaler man for aler my allas marlenarals hars you asfalfan in rat Sifalfaharis in stifan Valor anfinhans molyradar. - Talas Walls his farmer farend, vafarin, vern Holefanfar quallet renoven enerna, Nen fin 3nd New allysbranches rankfield las y wester

241

raussen im Schnee finde ich eine Henge von prachtvollen 1991schen illustrierten Worken, aber sie eind alle nass geseen und jetzt zu Eisklützen zusamangefroren...unrettbar erlorent

Dr. Hunder spricht mit einigen Spaniern, die hier im Quartie tehen, bittet, man möchte die Bücherole doch vermageln, und vor weiterer Veruneringung echtman. Unsern Zeit ist besetson, wir können selbet nicht sehr tuen, wir könnenzur berichten und verunlassen. Und das wird geschähen.

Auf uneers Veranlassung him ist sehr hald darnuf von sines ehren ehten-Sonierkomennde mit huwe die ganne erebischelische skils obset mach sign abgeholt worden und dort der Obsetz-Flete und Verwaltung des obersten orthodenen stelstichen mit rowen Hilden übergeben verden jeh hoffe, dans die spiter von rowen Hilden übergeben verden jeh hoffe, dans die spiter von

So ist schon naheau gans dunkel, whe wir des kwent verlasselven from to the des grosses Tores muss des Fahrer den Marses nocheel enhalten und die Notoriause aufrieden. Ich presiden nocheel enhalten und die Notoriause aufrieden ich so schon des Schnede, aus dem ich sie musst harvuchselen musstas die Schnede, aus dem ich sie musstas die Sille ist als goldig schimment auf; es ist eine kisine auf Sille ist als goldig schimment auf; es ist eine kisine sich der Sille ist eine Kannen versiorie, krousfunge Bromateliturdene, Richt elder aus ein und sie sille sich ein der der son auch der der son auch des der son und ver eine helbrunen lederene Bachdeniel aus de

WAS SEITDEM GESCHAH...

 Texterkennung für Typoskript mit nachtrainiertem Model ,einsatzfähig' (CER: 1,82% WER: 8,87% auf Evaluationsdatensatz)

- … Rechtslage leider unklar daher (noch) nicht publikationsfähig
- HTR auf Briefkorpus wegen Problemen bei Layout-Erkennung zunächst auf Eis gelegt (multi-direction layout recognition bislang nicht umsetzbar)

26 Smill van 4 Systember 1863.

Links Colis: - May hip is jo lains 24 Minister von grofa i is fall in Parfriel gelan, and sous Vago jo regamin Elligar Erifles Thanks, was as in dar hinkorns, mas jo nafranas ift. - Of fefor Januar nors Wajafor, rayorm is is Rinter la frift falls, is alla his not Hedrig befor mit reason. Coffers Jahan fis faiflest suffield not about Jeggiert, for Winterwolasten beganna, N. J. mil viballones jagisfell de If fall glinkling Maricas Topilge and sho Wagan bringen i foffer, right is is acron year Morals & Viletter nogerhrougen ratoman denait falls of the faccionspring no myarka mil star so freghan var Rieflar is the formers tister har Chychianalia man for aler my allas marlenands hers you asfalfan in rat Sifalfaharis in stifan Onfor antiphenan motyrafar. - ralas Walls fif favor favors, rafe itis, venan Hopfanfar quallet revoren enerna, shen fin 3nd New Whysbrander workfrifto las garafol

sid alle chne Feneterglas chne Turen. Durch Geschosseinschlüge und desclierte Sten und Winde sind die Wertvollen
Bücher mit Adsgelzsteinen, Schutt und Schnes bedeett. Tusen
dem eind diese Rüume als Abert benutzt werden gebet alleber
aus frühreren Jahrhunderten das mit Stige Burde und Lie ern
schlenen. Se erweist atch, dass beim Sinnug und Lie ern
die unserbiert kostbere Sibliothein noch in beste Ordnung par
die unserbiert kostbere Sibliothein noch in beste Ordnung par
Regalen und in Schrönken geordnet gewesen letzlie Spanlager
Regalen und ein Schrönken geordnet gewesen letzlie Spanlager loggifed und in Jenranson goordnet geween jet-vie Spanier haben aber allee was Mola war venheist und die Bücher piett les auf den Fuseboden geschittet. Tief erschüttert tetigen wir in den Rümen herun, such vir auf den netenhehn Bücherbergen stiefelnd, voreichtig um dicht in die Rümennete mu tersten. Die verben eine Auferbergen stiefelnd, voreichtig um dicht in die Rümennete mu tersten. Sie und die bestehn die Stenen die verben. ausende von Büchern eind aus dem 18ten, 17ten, vielleicht 6ten Jahrhundert.Gewiss ist es möglich, dass die Sowjets selbst schon einige Bestände rechtzeitig in Sicherheit ge-bracht haben oder dass Professor Sahm auch schon von hier einiges herausgesucht hat ... Zuch sollen, wie wir hören, der spanische Divisionsprediger, Offiziere und gebidete Soldaten apanische Livi monsprausger, wilniere und geriaere schaever der "Blauen Legion"dich de schom manches kostbere, gute nit genomen haben, sie funden neben ruseischen, deutschen fran-zögischen und englischen Herken viele lateinische und esgar apanische. Mir stellen aber auch japanische, chinestache, turkische, arabische und wer weise was für Schriften feet. Sin Raum enthült nur elawonische, vornehelich geietliche Litera-

braussen im Schnee finde ich eine Menge von pracht vollen ruseischen illustrierten Werken, aber sie sind alle nass ge-wesen und jetzt zu Eisklötzen zusamsengefroren...unrettbar

Dr. Wunder spricht mit einigen Spaniern, die hier in quartier schen, bittet, man wöchte die Büchervier doch vernacelen, und vor weiterer Verunschinkung echtisen. Zott ist beses-sen, wir können selbst nicht sehr mit Wönnennur berich-ten und verminssen. Und das wir tuen,

(Auf unsers Terminasum him ist sehr bald darauf von einem Hehrwachts-Somiestomminde mit LEN's die ganne ersbischbillebe siblichbek mach Miss abshebt worden und dert der Obbut, filege und eine hehre werden und dert der Obbut, filege und einem Minden der geschen werden, ich hoffe, das die spiter von den Sowjets wieder mach Howgored surfückgebracht worden ist!)

Se ist schon nahezu ganz dunkel, wie wir den Kreal verlassen "naittelber jeneelte des groosen Tores auss der Pahrez den Nagen nechsel anhalten und die Noteriaube aufheien. Ich spe-siere herne und ziche zwei starrgefrorme Saches und den Schnee, uns des ich die zwerz herundschen den dan den füllt alr els goldig schlassert auf es is den kleine, mit weldfranen verwierte, brouzförmige broat-altardecke. Bieht weit daven lugt eine hellbrauner ledernen Buchdeckel aus den Be ist schon nahezu ganz dunkel, wie wir den Kreml

WAS SEITDEM GESCHAH...

 Texterkennung für Typoskript mit nachtrainiertem Model ,einsatzfähig' (CER: 1,82% WER: 8,87% auf Evaluationsdatensatz)

• ... Rechtslage leider unklar daher (noch) nicht publikationsfähig

- HTR auf Briefkorpus wegen Problemen bei Layout-Erkennung zunächst auf Eis gelegt (multi-direction layout recognition bislangnicht umsetzbar)
- Drittes Einsatzszenario: (Vertriebenen-)Zeitungen
 - u. a. "Ostpreußenblatt" (1950-2003), "Preußische Allgemeine Zeitung" (2003-2023)
 - "Sudetendeutsche Zeitung" (derzeit: 1951-1955)
 - ... und viele andere mehr (noch nicht digitalisiert)
 - Aktuell vorliegend: > 3850 Ausgaben, > 83.000 Seiten

Gleiches Recht!

Text- und Layout Erkennung Abbyy FineReader, Sudetendeutsche Zeitung, Jg. 1 Folge 01, 07.04.1951

- - Lodgman bei Adenauer
- Lodgman bei Adenauer

- Layout wie erwartet schlecht zu erkennen
 - ... beste Ergebnisse bislang mit ABBYY FineReader und PERO OCR
 - ... auch hier allerdings Reading Order mitunter stark fehlerhaft

Text- und Layout Erkennung Abbyy FineReader, Sudetendeutsche Zeitung, Jg. 1 Folge 01, 07.04.1951



Lodgman bei Adenauer

- Layout wie erwartet schlecht zu erkennen
 - ... beste Ergebnisse bislang mit ABBYY FineReader und PERO OCR
 - ... auch hier allerdings Reading Order mitunter stark fehlerhaft
- Herausforderungen:
 - Erstellung Ground Truth sehr arbeitsaufwändig
 - Layout-Erkennungsmodelle von ABBYY und PERO ermöglichen kein Finetuning/Adaption?
 - Gleichzeitig hohe Anforderungen an Erkennungsqualität (v.a. Text, mittelbar Layout) durch Forschungsprojekt

FRAGEN

- Problem Layout-Erkennung; bei multi-direktionalen (Hand-)schriften, v. a. aber bei komplexen Zeitungs-Layouts.
- (Wie Layout-Erkennung evaluieren?)
- Tools für Layout-GT-Erstellung (eScriptorium zu umständlich für komplexe Layouts; LAREX? Weitere?)
- Publikation Evaluations-Datensätze? Verwendung als Trainingsmaterial nicht intendiert
- ... und was ist überhaupt realistisch zu erreichen/wo lohnt der eigene Arbeitsaufwand? → v. a. für Layout-Erkennung

Vielen Dank für die Aufmerksamkeit!

Ole Meiners Arbeitsbereich Forschungsdatenmanagement Abteilung Digitale Geschichte und Informationssysteme Herder-Institut für historische Ostmitteleuropaforschung