A-I-PoCoTo - Automatische und interaktive OCR Nachkorrektur

Klaus U. Schulz, Tobias Englmeier, Florian Fink

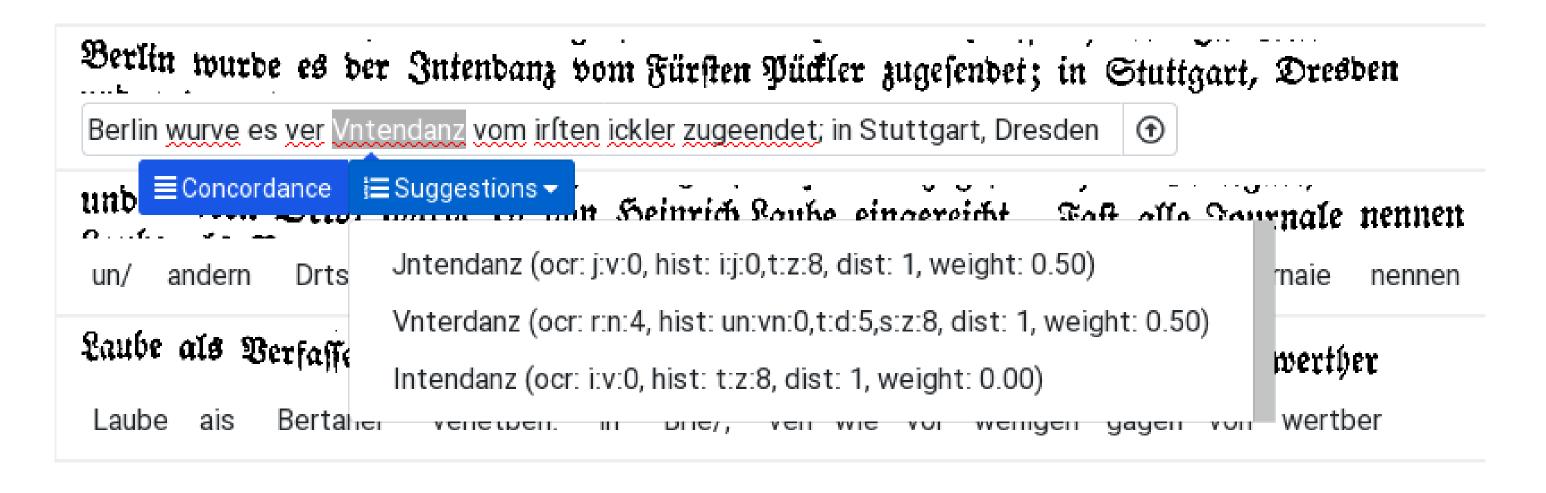
CIS - Centrum für Informations- und Sprachverarbeitung Ludwig- Maximilians- Universität München

A-PoCoTo (automatische Nachkorrektur)

Als Teil des OCR-D-Moduls 3 (Textoptimierung) wird in A-PoCoTo[1] eine vollautoma-tische Nachkorrektur OCR-erkannter historischer Dokumente umgesetzt. Ein besonderes Augenmerk wird dabei sowohl auf die Behandlung historischer Rechtsschreibvarianten als auch auf die Verhinderung von Verschlimmbesserungen gelegt.

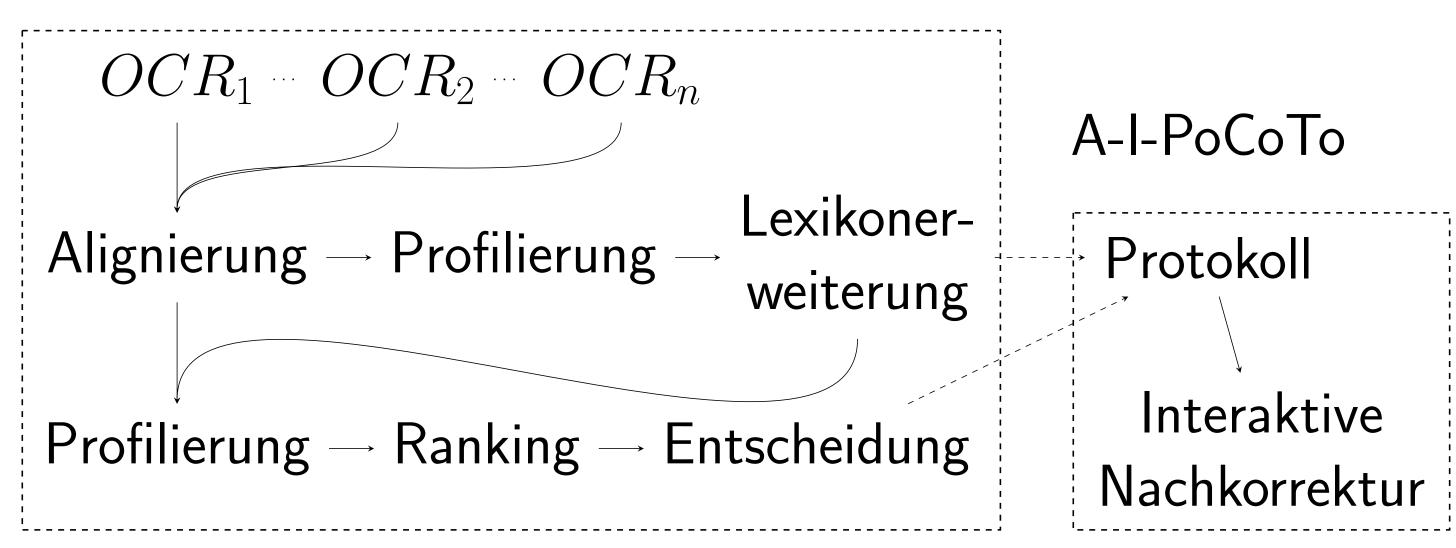
Sprachprofilierung der OCR-Ausgabe

Die am CIS entwickelte *Profilierungs*-Technologie[3, 4] leitet eine statistische Abschätzung über ein OCR-erkanntes historisches Dokument auf der Basis unterschiedlicher Hintergrundlexika und einer Menge *historischer Patterns* (wie in etwa $lei \rightarrow ley$ oder $t \rightarrow th$) ab und erzeugt für jedes OCR-Wort gewichtete *Interpretationen* der Form $w_{mod} \rightarrow_{\alpha} w_{hist} \rightarrow_{\beta} w_{ocr}$.



Arbeitsweise

A-PoCoTo



Die automatische Nachkorrektur läuft über drei (zwei) Klassifikationsschritte: die optionale Lexikonerweiterung (Extraktion unbekannter Lexikoneinträge zur Erweiterung des Profilierungslexikons), den Ranker (Neusortierung der Profilerinterpretationen) und den Entscheider (Enscheidung über die Ausführung von Korrekturen), mit zwei (einem) zwischengeschalteten Profilierungsdurchläufen. Als Eingabe dienen hierbei eine Haupt-OCR-Erkennung (OCR_1) sowie n-1 weitere Hilfs-OCR-Erkennungen $(OCR_2...OCR_n)$. Die Lexikonerweiterungen sowie die durchgeführten Korrekturen werden protokolliert und dienen der späteren Weiterverarbeitung durch A-I-PoCoTo (siehe rechte Spalte).

Das Training geht dokumentenweise vor, wobei für jedes Trainingsdokument ein eigenes Profil erstellt wird. Für jeden Schritt werden $1\dots n$ Klassifikatoren mit einer entsprechenden Anzahl von Eingangs-OCR's trainiert. Jedem Klassifikator wird dabei flexibel eine Feature-Menge zugewiesen, die beim Training (und auch bei der Nachkorrektur) je nach vorhandenen OCR-Eingaben dynamisch an- und abgeschaltet werden.

Auswertungen

	mit Lexil 1 OCR	konerweiterur 2 OCR	ng ohne Lexi 1 OCR	konerweiterung 2 OCR
Korrekte OCR Token	1891	1891	1891	1891
Inkorrekte OCR Token	1007	1007	1007	1007
Keine korrekten Vorschläge	720	917	718	718
Top 1 Vorschlag korrekt (Profiler)	128	42	128	128
Top 1 Vorschlag korrekt (Ranker)	219	124	218	223
Erfolgreiche Korrekturen	207	53	206	141
Verpasste Korrekturen	12	34	12	82
Unglückliche Korrekturen	539	71	508	12
Korrekte Token nach der Korrektur	1560	1985	1589	2020

[&]quot;1557, Bodenstein, WieSichMeniglich" [1].

A-I-PoCoTo (interaktive Nachkorrektur)

Es ist nicht davon auszugehen, dass mit einer vollautomatischen Korrektur stets die gewünschte Akkuratheit eines Texts erreicht wird. Aus diesem Grund wird mit A-I-PoCoTo eine interaktive Nachkorrektur ermöglicht[5]. Sowohl die Lexikonerweiterung als auch die vorgenommenen automatischen Korrekturen können in A-I-PoCoTo manuell überprüft und entweder bestätigt oder abgelehnt werden.

Wie bereits bei der *adaptiven Profilierung*[2] können in weiterführenden Arbeiten bei der interaktiven Nachkorrektur sowohl die verwendeten OCR-Modelle als auch die Modelle der automatischen Nachkorrektur unter Berücksichtigung der manuellen Korrekturen weiter verfeinert werden.

Konkordanz

○ Toggle	e selection	Correction	1 Set correction	→ Correct selected
33:25	Hanns	Christoff Tanner Hof-Oberrichter Ambtsverwalter, hristoff Tanner Hof-Oberrichter Ambtsverwalter, Ambtsvervslter,		>
41:28		Adolph von Peffenhausen, Mundtschenckh, auf Ine Adolph von Poffonhausen, Mundtschenelilh, auf Ine		>
45:15	Hanns	Cramer Hofrathsdiener Soldt 28 fl. 34 kr. 2 hl., vnnd Crumor Hofrathsdienr Snldt s I. I Er. Ehl. vnnd		>
51:24		Frelich verwalter der Comedj Claider Soldt vnd		>

Interaktive Lexikonerweiterung

	Extensions			Unknown
	Search:			Search:
Word	↑↓ Frequency	11	Word	↑ Frequency
dieſer	111	>	< li>fich	424
dieſe	102	>	< ift	343
deutschen	82	>	< daß	201
deutschland	62	>	< find	102
geſchichte	34	>	< fehen	25

Interaktive Nachkorrektur

	Always		Sometimes		Never
	Search:	Sea	arch:	S	Search:
Word ↑	Frequency	↑] Word ↑]	Frequency	Word ↑↓	Frequency
canlor	Canlor → Canlos ⊘		Horr → Horr ⊘	adolplh	Adolplh → Adolpph ⊗
canxler	Canxler → Cancler ⊘	horrn	Horrn → Horrn ⊗	aelhigen	aelhigen → aezhigen ⊗
derliehen	derliehen → uerliehen ⊘		Horrn → Horrn ⊗ Horrn → Horrn ⊗	ainon	ainon → ainotn ⊗
dero	Dero → Derp ⊘		Horrn → Horrn ⊗	ambt	Ambt → Ambt ⊗
dionor	Dionor → Sonor ⊘	iohann	Iohann → Iohann ⊗	aurelio	Aurelio → Aurels ⊗
doder	doder → doder ⊘		Iohann → Iohann ⊘	bahier	bahier → bahzer ⊗
edem	edem → ekzem ⊘	liferung	Liferung → Liferung ⊗ Liferung → Liferung ⊗	bayrn	Bayrn → Bayrz ⊗
ehorr	EHorr → EHer ⊘	vnnd	vnnd → pvnnd ⊗	bifl	bifl → bifz ⊗
entrieht	entrieht → entriemt ⊘	VIIIIG	Vnnd → Vnnd ⊘	bornhnrd	Bornhnrd → Bernhard ⊗
erou	erou → eroz ⊘		vnnd → pvnnd ⊗	ciidt	Ciidt → Ciedt ⊗

Referenzen

- [1] Tobias Englmeier, Florian Fink, and Klaus U. Schulz. A-I-PoCoTo: Combining Automated and Interactive OCR Postcorrection. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, DATeCH2019, page 19–24, New York, NY, USA, 2019. Association for Computing Machinery.
- [2] Uwe Springmann Florian Fink, Klaus U. Schulz. Profiling of ocr'ed historical texts revisited. In *Proc. 2nd Conference Digital Access to Textual Cultural Heritage (DATeCH 2017)*, pages 61–66. ACM Digital Library, 2017.
- [3] Ulrich Reffle. Efficiently generating correction suggestions for garbled tokens of historical language. Natural Language Engi-
- neering, 17(2):265-282, 2011.
- [4] Ulrich Reffle and Christoph Ringlstetter. Unsupervised profiling of OCRed historical documents. *Pattern Recognition*, 46(5):1346–1357, 2013.
- [5] Thorsten Vobl, Annette Gotscharek, Uli Reffle, Christoph Ringlstetter, and Klaus U. Schulz. PoCoTo an Open Source System for Efficient Interactive Postcorrection of OCRed Historical Texts. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, DATeCH '14, pages 57–61, New York, NY, USA, 2014. ACM.