

OCR-Evaluierung von Artikelüberschriften mit (lokalen) LLMs

Eine Evaluierung

Worum geht es - Problematik

- In unseren Zeitungsprojekten mit Artikelsegmentierung wird häufig eine hohe OCR-Qualität der Artikel-Überschriften verlangt.
- Diese kann nur durch manuelle Nachprüfung/Nachkorrektur erreicht werden.
- Teilweise ist das nur durch „double-keying“ erreichbar.
- Der manuelle Aufwand dieser Nachkorrektur ist ein signifikanter Kostenfaktor im Projekt.

Worum geht es - Idee

- Nutze (lokale) LLMs zur Bewertung der OCR
 - Wir bevorzugen lokale LLMs aus Kostengründen – nehmen ChatGPT gerne als qualitativ hochwertige Referenz
 - Nutze diese Bewertung, um die Menge der Artikelüberschriften, die zur manuellen Nachkorrektur vorgelegt werden, zu reduzieren.
- ➔ Dieses ist eine Weiterentwicklung der Idee vom letzten Jahr:
„Automatische OCR-Korrektur mit LLMs“:
Dort haben wir versucht, die OCR von kompletten Artikeltexten durch LLMs automatisch zu korrigieren
- ➔ Diese neue Idee hat geringere Komplexität bei hoher kommerzieller Relevanz

Unser Ansatz im Überblick

- Nimm des OCR-Ergebnis einer Artikelüberschrift und frage das LLM
- Frage 1: Korrigiere OCR
 - Tue das mit mehreren LLMs
 - Vergleiche Mehrheitsvoting mit OCR-Result
 - Wenn nicht gleich, dann „bad“
- Frage 2: Frage nach einer Confidence (0..1) von der OCR
 - Tue das mit mehreren LLMs
 - Wenn Minimum kleiner als Schwellwert (heuristisch bei uns <0.3), dann „bad“

Unser Ansatz - Details

- Ausprobierte LLMs:
 - qwen, phi, mistral (mit verschiedenen Quantisierungen)
- Übrig gebliebene, benutzte LLMs
 - qwen mit verschiedenen Quantisierungen
 - Die anderen haben keine neue Informationen gebracht (im Sinne von neuer „bad“ Erkennung)

Prompts – Frage 1 - Beispiele

You are an OCR correction engine.

Output only the corrected headline

no quotes, no explanations, no labels.

If nothing needs correcting, output the input exactly.

You are an OCR correction engine.

Input is a newspaper headline.

Correct OCR errors only (characters, spacing, obvious misspellings).

Do not rephrase or change style/capitalization.

Output only the corrected headline—no quotes, no explanations, no labels.

If nothing needs correcting, output the input exactly

Prompts – Frage 2 - Beispiele

You are an OCR confidence estimation engine.

Input is a newspaper headline.

Output only a single number between 0 and 1 indicating the confidence that the headline is correct.

0 means definitely incorrect, 1 means definitely correct.

You assess the OCR result of a newspaper headline.

Output only a single number between 0 and 1 indicating the confidence that the OCR is correct.

0 means definitely incorrect, 1 means definitely correct.

Technische Details

- Benutzte Software/Hardware

- Olama-Framework
- GPU: P100 (16 GB VRAM)

- Laufzeiten

- Frage 1: Ca. 1.2 Headlines/s
- Frage 2: Ca. 0.2 Headlines/s (6x langsamer!)

➔ Varianten von Frage 2, nicht nach quantitativen Zahlen zu fragen, sondern nach qualitativer Bewertung („Ist das gute/schlechte OCR“) ist viel schneller (wie Frage 1), aber im Ergebnis viel schlechter (nicht brauchbar)

Evaluierungsdat

- 1014 Headlines der New York Post Ausgaben aus 1998
- Somit modern, eher schwieriges Layout
- Absichtlich gewählt – OCR ist delementsprechend auch nicht so gut



Ergebnisse - I

		correction_confidence assessment			
		good_good	good_bad	bad_good	bad_bad
OCR					
good	919 (90.6%)	595 (58.7%)	117 (11.5%)	145 (14.3%)	62 (6.1%)
bad	95 (9.4%)	10 (1.0%)	4 (0.4%)	27 (2.7%)	54 (5.3%)

Total headlines: 1014

Correction&Confidence: missing 10 (1.0%) headlines by correcting only 409 (40.3%)
 Correction only: missing 14 (1.4%) headlines by correcting only 288 (28.4%)

Ergebnisse - II

Character **Error** Rate Distribution vs AI Assessment for ocr_result

AI Assessment	Count	of errors	of total chars
good_good	11	0.37%	0.04%
good_bad	299	10.02%	1.19%
bad_good	679	22.75%	2.70%
bad_bad	1996	66.87%	7.92%

Ergebnisse - III

Total Chars:	25202	Chars
Total Headlines:	1014	Headlines
Average Char/Headline	24,9	Chars/HL

good_good-Case		
not good_good headlines:	409	Headlines
Estimated Char for not good_good:	10165	Chars
Errors from good_good (missing)	11	Errors/Lev
Average manual correction error rate	1%	
Error from manual correction of not good_good	102	Errors/Lev
Total CER	0,45%	

good_*-Case		
not good_* headlines:	288	Headlines
Estimated Char for not good_good:	7158	Chars
Errors from good_good (missing)	310	Errors/Lev
Average manual correction error rate	1%	
Error from manual correction of not good_good	72	Errors/Lev
Total CER	1,51%	

Ergebnisse - IV

- CER von
 - OCR: 11,85%
 - Nach manueller Korrektur: 1,25%
 - KI-Vorschlag: 21,45%
- ➔ Neben-Ergebnis:
 - Manuelle Korrektur erreicht 1.25% CER
 - Pauschale Nutzung des OCR-Korrektur-Vorschlags der KI als neues OCR-Ergebnis ist NICHT hilfreich!

Bespiele für „false positiv“ (missing) good_good



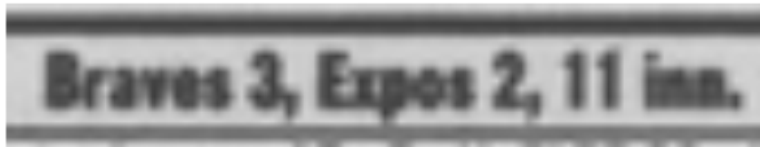
Cowboys coach
Barry Switzer quits

Cowboys coach
Barry Switzer quits



Music

Music



Braves 3, Expos 2, 11 inn

Braves 3, Expos 2, 11 inn.

Bespiele für „false positiv“ (missing) good_bad



NO The Brooklyn Museum turned down a possible deal with the mayor yesterday and decided instead to sue the city over plans to slash public funds.
DUNG DEAL

NO
DUNG DEAL



Page
Six.

Page
SixTM



STRANDED
MOM GETTING
JUMP-START
ON NEW LIFE

STRANDED
MOM GETTING
JUMP-\$TART
ON NEW LIFE

Fazit

- Dieser Ansatz ist für uns vielversprechend
 - Mit dem Ziel, die Kosten für die manuelle Nacharbeit zu reduzieren
 - So wird bei der Variante „Frage 1“ & „Frage 2“ (good_good), eine Korrekturmengenreduktion um 60% erreicht.
 - Dabei bleiben 10 falsche Headlines ungeprüft, das sind aber nur 0.4% aller falschen Zeichen.
 - Zusätzlich ist (wahrscheinlich) die Gesamt-CER reduziert gegenüber dem rein manuellen Ansatz.
- Nächster Schritt: Erweiterung der Test-Datenbasis



Kontakt

CCS Content Conversion Specialists – Hamburg/Germany

Stefan von der Heide (CTO)

Stefan.vonderHeide@content-conversion.com

Telefon: +49 1579 2366592

info@content-conversion.com

[www. content-conversion.com](http://www.content-conversion.com)