Koninklijke Bibliotheek
National Library of the Netherlands

Ochre, a Toolbox for OCR Post-Correction

Ochre is an open source software package that allows users to perform different kinds of tasks related to OCR post-correction, including:

- Training character-based language models using LSTMs
- Assessing post-correction performance
- Word level error analysis
- Preprocessing existing data sets

Download ochre from

https://github.com/KBNLresearch/ochre

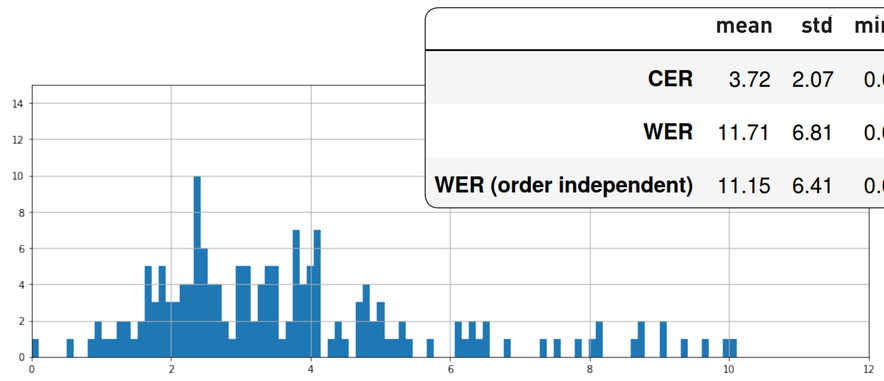
Ochre was tested on the VU DNC Corpus

- 3340 newspaper articles (ocr text and gold standard)
- Articles with different new genres from 5 Dutch national newspapers (1950/1951)
- Noisy!
- 5% validation set, 5% test set
- http://dev.clarin.nl/node/4194

Post-correction performance

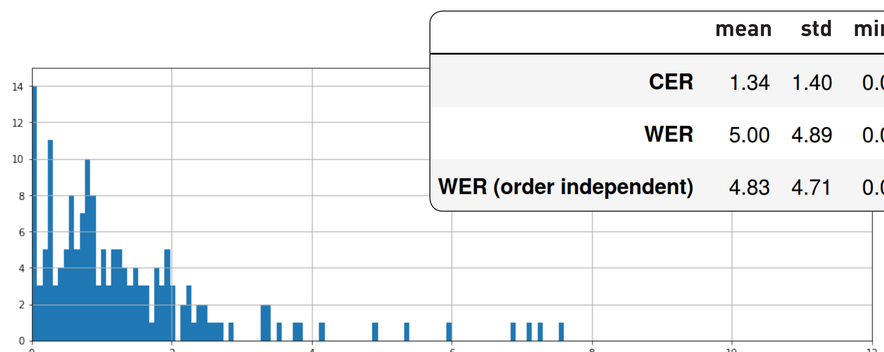
- Neural network architecture: seq2seq (2x256 nodes)
- Input: sequence of 25 characters ocr text
- Output: sequence of 25 characters gold standard text

OCR text vs. gold standard



	mean	std	min	max
CER	3.72	2.07	0.0	10.11
WER	11.71	6.81	0.0	45.00
WER (order independent)	11.15	6.41	0.0	45.00

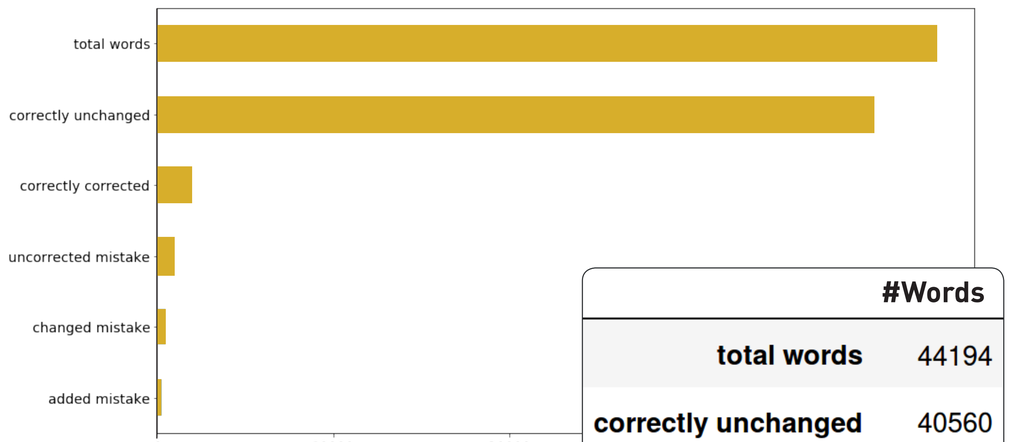
Post-correced using seq2seq vs. gold standard



	mean	std	min	max
CER	1.34	1.40	0.0	7.60
WER	5.00	4.89	0.0	28.33
WER (order independent)	4.83	4.71	0.0	28.33

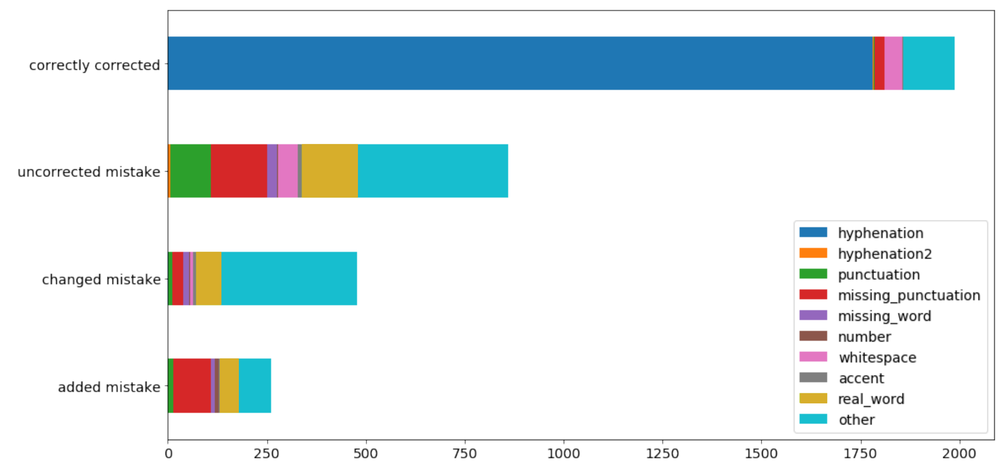
Word-level Error Analysis (1)

- Given OCR text, post-corrected OCR text and gold standard text, ochre counts the words that should and shouldn't be corrected.
- The test set contains 3365 words with OCR mistakes,
- Most of which are corrected!



	#Words
total words	44194
correctly unchanged	40560
correctly corrected	1999
uncorrected mistake	869
changed mistake	497
added mistake	269

Word-level Error Analysis (2)



- Ochre classifies the differences between words into error types
- Unfortunately, the corrected errors are mostly hyphenation errors...
- Other errors do not occur often enough. Future work: generate training data to learn to correct these mistakes.

Janneke van der Zwaan
Netherlands eScience Center
j.vanderzwaan@esciencecenter.nl

Lotte Wilms
Koninklijke Bibliotheek
lotte.wilms@kb.nl

netherlands eScience center