OCR Quality and NLP Preprocessing

Margot Mieskes **Stefan Schmunk**
Darmstadt University of Applied Sciences
firstname.lastname@h-da.de

Abstract

We present initial experiments to evaluate the performance of tasks such as Part of Speech Tagging on data corrupted by Optical Character Recognition (OCR). Our results, based on English and German data, using artificial experiments as well as initial real OCRed data indicate that already a small drop in OCR quality considerably increases the error rates, which would have a significant impact on subsequent processing steps.

1 Introduction

Humanities are going more and more digital, which makes it more important for researchers in the Humanities to simply use tools and get reasonable results. An NLP specialist is not always at hand to support them in the technical aspects of answering their research questions. Additionally, with the rise in the publicity of what NLP tools can achieve, the expectations rise that these tools can be used out of the box. Often in the Humanities, the material on which the research is based on is only scanned and ran through Optical Character Recognition (OCR) tools. Depending on the quality, this might have a considerable and potentially significant impact on the quality and therefore the results of various NLP tools. The open question in this context is, how much do basic NLP tasks degrade with the increase in OCR errors. We give preliminary answers to this question based on various corpora, using Part of Speech (POS) tagging. Our results indicate that even with fairly low OCR errors, which are currently only obtained under ideal circumstances, the performance of standard off-the-shelf tools drops considerably.

2 Related Work

Work on explicitly studying POS tagging is vast, but there is very little on tagging historical data. Yang and Eisenstein (2016) look into POS tagging of historical English data. They point out that accuracy for taggers drops from 97% for the British National Corpus to 82% for Early Modern English. Their results show that using temporal adaptation a 30% relative error reduction can be achieved. Hardmeier (2016) presents a character-level neural network for POS tagging of historical texts. The author evaluates the performance on Swedish verb identification and German POS tagging, achieving an F-score of 0.85 on the test set for Swedish and 0.86 on the German test data. But a big issue when processing historical data is the digitization quality and more over the quality of the optical character recognition (OCR). Strange et al. (2014) describe common problems such as "smudged, faded and warped text". They also report that the accuracy levels of OCRed newspapers vary considerably and an accuracy of 94.5% can be achieved under almost ideal circumstances, going down to 65%. Kettunen and Pääkkönen (2016) analysed the digitization quality also with respect to the decades the data was published. Their results range from 60% in 1770–1779 up to 70% in 1900–1910, with the highest level of 73.7% for data from 1880–1889. Starting in 2013, the German Research Foundation (DFG) funded a series of library pilot projects that evaluated the challenges of digitizing newspapers. So far, however, there are no valid studies that show what quality level of OCR is needed for different procedures and different domains (Klaffki et al., 2018).

0.05 CER	0.10 CER	0.15 CER	0.20 CER	0.25 CER
sie	sie	sie	SolOhergestalt	molchergestaYt
beC	beC	beC	istgsie	isr
naho	naho	naho	AeiNnphe	sQe
drittjhalb-hundert	drittjhalb-hundert	drittjhalb-hundert	driWtehalb-hsn0ert	bViJflhe
Oahg	Oahg	Oahg	Jahinunper	driCtYhalb-5unds4U
unter2lauter	unter2lauter	unter2lauter	Eax8er	JJhr
Biastischen	Biastischen	Biastischen	BiastisEhen	ukter
Fwrsten	Fwrsten	Fwrsten	Ckrsten	laYter
aufg7wachsen	aufg7wachsen	aufg7wachsen	qufgewachseCIc	Si4stitcheG
.	.	.		FürsGenwaPfgewachsen6HS

Table 1: An example from Zeller Chronik02 1738 with artificially added 5%, 10%, 20% and 25% Character Error Rate.

	WSJ					Brown					DTA				
F_1	0.05	0.10	0.15	0.20	0.25	0.05	0.10	0.15	0.20	0.25	0.05	0.10	0.15	0.20	0.25
	0.96	0.93	0.90	0.88	0.84	0.93	0.93	0.93	0.93	0.93	0.74	0.73	0.73	0.72	0.71

Table 2: Results for the WSJ, the Brown Corpus and the DTA with artificially added CER.

3 Experiments and Results

For the implementation of our experiments we use Python 3 and NLTK and rely only on off-the shelf tools and implementation, for example for taggers and evaluation metrics. Evaluation is based on precision, recall and F-measure relying on the implementation provided by the NLTK metrics package.

Data We use three different data sets: The *Brown Corpus* (Francis, 1965) and the *Wall Street Journal* (WSJ) (Paul and Baker, 1992), both as available in the NLTK data package. The third corpus is the *Deutsche Textarchiv* (DTA), which contains German documents from the 17^{th} to the early 20^{th} century (Geyken et al., 2018), covering 1500 documents with approx. 120 million words, also containing linguistic annotations (Bański et al., 2018). Additionally, various genres are represented such as scientific literature and newspapers, but also biographies, letters and fictional work, such as plays or fairy tales. The data containing the linguistic annotations has been made publicly available in February 2019 and contains 4410 texts.[1] From this corpus, we use 22 documents with approx. 1 million tokens and slightly over 50.000 sentences.[2] Therefore, in terms of size it is comparable to the two English corpora.

Lab Experiment OCR errors can range between 5% under ideal circumstances (Strange et al., 2014) and up to 40% for documents from the 18th century (Kettunen and Pääkkönen, 2016). Therefore, we randomly introduced errors to the documents in our experiments by replacing 5%, 10%, 15%, 20% and 25% of the characters with random other characters such as upper- and lower-case letters and digits. A qualitative analysis of the artificially changed data reveals that our random insertion of errors reflect errors in real OCR data: words get concatenated, characters are exchanged and full stops get lost (see Table 1).

OCR Experiment In addition to the laboratory experiments, we randomly selected several pages from the DTA documents that were also part of our laboratory experiment, to run them through Abbyy Finereader Online[3] in a standard configuration.

Results With only 5% CER results are at $F_1 = 0.96$ and drop to $F_1 = 0.84$ for 25% CER. For the Brown Corpus results drop to approx. $F_1 \geq 0.43$. The standard deviation is small, which indicates that documents from individual domains do not perform very differently and results do not depend on the domain (see also Table 3). The Brown Corpus data seems to be more affected by errors than the WSJ. One reason might be, that the standard NLTK tagger is trained on WSJ data and more easily recovers errors. For both data sets the decrease in sentences and tokens is roughly proportional to the increase in CER: with 25% CER the number of sentences in the WSJ drops by 1/3 and the number of tokens by 1/4, which also affects POS tagging quality.

The results for the DTA are similar to the two previous result sets and range in between them. The

[1] `http://media.dwds.de/dta/download/dta-lingattr-tei_2019-02-06.zip`

[2] Due to time constraints we could not process the whole set, but plan to do so in the near future.

[3] `https://finereaderonline.com/en-us`

	simple tag set					full tag set				
Genre	0.05	0.10	0.15	0.20	0.25	0.05	0.10	0.15	0.20	0.25
Science Fiction	0.93	0.93	0.93	0.93	0.93	0.44	0.45	0.44	0.44	0.43
Lore	0.93	0.93	0.93	0.93	0.93	0.45	0.47	0.47	0.46	0.45
Learned	0.93	0.93	0.93	0.93	0.93	0.49	0.48	0.47	0.47	0.47
Reviews	0.95	0.95	0.95	0.95	0.95	0.41	0.43	0.42	0.42	0.41
Humor	0.94	0.94	0.94	0.94	0.94	0.43	0.44	0.44	0.43	0.43
News	0.93	0.93	0.93	0.93	0.93	0.42	0.44	0.43	0.42	0.42
Government	0.92	0.92	0.92	0.92	0.92	0.44	0.45	0.45	0.44	0.43
Mystery	0.92	0.92	0.92	0.93	0.92	0.43	0.44	0.44	0.43	0.43
Fiction	0.93	0.93	0.93	0.93	0.93	0.44	0.45	0.44	0.44	0.43
Editorial	0.93	0.93	0.93	0.93	0.93	0.41	0.42	0.42	0.42	0.41
Belles Lettres	0.93	0.93	0.93	0.93	0.93	0.45	0.47	0.46	0.47	0.45
Adventure	0.93	0.93	0.93	0.93	0.93	0.45	0.44	0.45	0.44	0.44

Table 3: Results for the Brown Corpus sorted by Genre.

standard deviation is fairly low, which indicates that individual documents from various centuries do not deviate from these results. We ran four documents through the Abby Online service. These preliminary experiments achieve F_1 scores ranging from 0.63 up to 0.70. This data set is too small, to determine whether this difference is statistically significant, but they are roughly 10% points (absolute) below the results for the artificial settings. We observe that the quality of the material is very poor, which impacts the OCR quality, most likely resulting in CER considerably higher than in our artificial experiments. But, it also shows that our experimental results obtained from the laboratory experiments allow for tentative conclusions with respect to OCR quality required for further analysis using NLP tools.

4 Discussion and Conclusion

We performed two sets of experiments on English and German data, artificially introducing OCR errors to clean data and evaluating the performance of POS taggers on this data. Initial experiments using an off-the-shelf OCR tool, without any further parameter tweaking, indicate that the results obtained through the artificial settings allow for conclusions with respect to the OCR quality necessary to use NLP tools. Already 5% CER have a huge impact on the performance of the POS taggers both on the English and the German data. Only on a data set, which was used to train the POS tagger the results are competitive at 5% CER, but drop considerably with higher error rates, which are usually observed in automatic OCR. These results indicate that researchers in the Humanities still need technical expertise in using and tweaking the tools at hand and the NLP community should continue their efforts to provide easy to use and reliable tools for processing digitized, historical data.

Future Work There are a range of open questions to be addressed in this scenario. In the short-term a more in-depth analysis of the tagging errors with respect to specific tags could reveal ways to improve the tagging results without having to perform a complete manual annotation. Results on the WSJ indicate, that re-training an existing tagger would improve the performance on the historical data. We also plan to look into more details about the error distribution with respect to time period and genre.In a next step, 10-year samples from the directories of German-language prints of the 16th, 17th and 18th centuries, which are already digitized are to be created. Some of these collections are available as digitized documents, but they are still largely not OCRed. As a result, our analytical approach can significantly contribute to the quality grades in which this material must be developed in order to generate machine-readable, annotatable and processable data. In the longer-term improving the OCR to deal with historic spelling and printing could boost the accuracy considerably. Additionally, it would be interesting to perform an extrinsic evaluation of the improved tools and procedures by evaluating them as part of a pipeline for Named Entity Recognition or Event Detection to analyse effects of various error sources with respect to other NLP tasks.

Acknowledgements

This work has been supported by the research center for Digital Communication and Media Innovation (DKMI) and the Institute for Communication and Media (IKUM) at the University of Applied Sciences Darmstadt.

References

Piotr Bański, Susanne Haaf, and Martin Mueller. 2018. Lightweight Grammatical Annotation in the TEI: New Perspectives. In *Proceedings of the 11th Conference on Language Resources and Evaluation*, pages 1795–1802, Miyazaki, Japan, May.

Antske Fokkens, Serge ter Braake, Niels Ockeloen, Piek Vossen, Susan Legene, and Guus Schreiber. 2014. BiographyNet: Methodological issues when NLP supports historical research. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland.

W. Nelson Francis. 1965. A Standard Corpus of Edited Present-Day American English. *College English*, 26(4):267–273.

Alexander Geyken, Matthias Boenig, Susanne Haaf, Bryan Jurish, Christian Thomas, and Frank Wiegand. 2018. Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN. In Henning Lobin, Roman Schneider, and Andreas Witt, editors, *Digitale Infrastrukturen fr die germanistische Forschung*, pages 219–248. De Gruyter.

Christian Hardmeier. 2016. A neural model for part-of-speech tagging in historical texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 922–931. The COLING 2016 Organizing Committee.

Kimmo Kettunen and Tuula Pääkkönen. 2016. Measuring lexical quality of a historical finnish newspaper collection – analysis of garbled ocr data with basic language technology tools and means. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Lisa Klaffki, Stefan Schmunk, and Thomas Stäcker. 2018. Stand der Kulturgutdigitalisierung in Deutschland. Technical report, Georg-August Universität Göttingen. Dariah-DE Working Paper Nr. 26.

Douglas B. Paul and Janet M. Baker. 1992. The design for the wall street journal-based csr corpus. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '91, pages 357–362, Stroudsburg, PA, USA. Association for Computational Linguistics.

Carolyn Strange, Daniel McNamara, Josh Wodak, and Ian Wood. 2014. Mining the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers. *Digital Humanities Quarterly*, 8(1).

Yi Yang and Jacob Eisenstein. 2016. Part-of-speech tagging for historical english. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1328. Association for Computational Linguistics.