Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)

BERT for Named Entity Recognition in Contemporary and Historical German

Kai Labusch
Staatsbibliothek zu Berlin - Preußischer Kulturbesitz
10785 Berlin, Germany
kai.labusch@sbb.spk-berlin.de

Clemens Neudecker
Staatsbibliothek zu Berlin - Preußischer Kulturbesitz
10785 Berlin, Germany
clemens.neudecker@sbb.spk-berlin.de

David Zellhöfer
Staatsbibliothek zu Berlin - Preußischer Kulturbesitz
10785 Berlin, Germany
david.zellhoefer@sbb.spk-berlin.de

Abstract

We apply a pre-trained transformer based representational language model, i.e. BERT (Devlin et al., 2018), to named entity recognition (NER) in contemporary and historical German text and observe state of the art performance for both text categories. We further improve the recognition performance for historical German by unsupervised pre-training on a large corpus of historical German texts of the Berlin State Library and show that best performance for historical German is obtained by unsupervised pre-training on historical German plus supervised pre-training with contemporary NER ground-truth.

1 Introduction

The transformer (Vaswani et al., 2017) is a recent neural network architecture that has been used as the central building block of representational language models such as GPT (Radford et al., 2018) or BERT (Devlin et al., 2018). These representational models can either be utilized to derive features that serve as input for other models such as a long short term memory (LSTM) and/or a conditional random field (CRF) or they can be directly trained on some supervised task. In this paper, we follow the latter approach and train a pre-trained BERT model directly for named entity recognition (NER) tasks.

In contrast to contemporary German, historical German texts pose multiple challenges on a potential algorithm because their language is less standardized and their digital representation has been typically obtained by optical character recognition (OCR) that has been shown to be error prone in this particular scenario (Federbusch et al., 2013).

In the experiments presented below, we evaluate the performance of BERT on two contemporary German NER data sets as well as on three different historical German NER corpora (see Sec. 5). We get best results for historical German by application of unsupervised pre-training on a large historic german text corpus plus supervised pre-training using contemporary German NER ground-truth. In contrast best results for contemporary German are obtained without unsupervised pre-training. The large historical German text corpus that is used for unsupervised pre-training has been extracted from the digital collections of the Berlin State Library (Staatsbibliothek zu Berlin/SBB).

The software used in the experiments is provided for download [1].

2 Background

The SBB is digitizing its copyright-free holdings and makes them publicly available online in various formats for direct[2] or automated[3] download. As part of an on-going process, a growing amount of OCR-derived full-texts of the digitized printed material is provided in ALTO[4] format but is mainly used for internal use cases such as full-text indexing and other information retrieval tasks.

However, OCR of historic documents is significantly more difficult than OCR of modern texts due to the large variety of fonts, layouts, mixed languages, and non-standardized orthography of printed texts from before 1850. As a consequence, texts generated by standard OCR contain a high amount of word errors. Similar challenges have been described by (Lopresti, 2009) and (Alex and Burns, 2014) who have noted that the quality of text analysis is directly tied to the level of noise in a document. Additional difficulties are caused by the historic language (Piotrowski, 2012).

Despite these obstacles, natural language processing – and NER in particular – strongly con-

[1] https://github.com/qurator-spk/sbb_ner
[2] https://digital.staatsbibliothek-berlin.de
[3] https://digital.staatsbibliothek-berlin.de/oai
[4] https://www.loc.gov/standards/alto/

tribute to an improvement of the user experience as they leverage supportive means for exploration and search within large text corpora. Furthermore, a growing research interest from the Digital Humanities in text and, e.g., data mining for historical social network analysis relies on the extraction of named entities from the digitized and OCR-derived full-text collections.

First experiences with NER for historical texts at the SBB were obtained in the Europeana Newspapers project where a CRF (Finkel et al., 2005) was trained on manually labeled OCR texts of historic newspapers (Neudecker et al., 2014). This approach was superseded in the Oceanic Exchanges project where (Riedl and Padó, 2018) achieve state of the art results for historic German by combining a bidirectional long short term memory (biLSTM) with a CRF as top layer and transfer learning.

The work presented in this paper aims towards a versatile approach that performs decently on texts of different epochs, i.e. contemporary and historical, without requirement of intense parameter tuning with respect to particular target corpora.

The paper is structured as follows: The next section outlines the relevant work in the context of the presented approach. Section 4 describes four data sets that are used in the three experiments presented. In particular, it presents the data of the Berlin State Library that has not been published so far. Then, Section 5 gives a brief description of the technical details of the experiments. The outcome of the experiments is discussed and interpreted in Section 6. The paper concludes with an outlook on future work.

3 Related Work

(Grover et al., 2008) designed a rule-based system for recognizing person and place names in digitized records of British parliamentary proceedings from the late 17th and early 19th centuries and report F_1–scores from 70.35 to 76.94 percent.

(Packer et al., 2010) compare the performance of a dictionary-based extractor, a regular expression rule-based extractor, a Maximum Entropy Markov Model (MEMM) and a CRF on historical OCR-processed documents with a mean word error rate of 56 percent, revealing that a voting-based ensemble method can boost F_1–scores from 60.7 to 68 percent.

For a corpus of historic French newspapers, (Gaibert et al., 2012) report F_1–scores between 55.2 and 68.9 percent for two stochastic and one rule-based system by including noisy entities in the annotations.

In the Europeana Newspapers project, (Neudecker et al., 2014) measure F_1–scores of 46.6 to 73.27 percent with a CRF trained on annotated noisy OCR from historic newspapers in Dutch, French, and German. F_1–scores up to 60 percent are obtained for a dataset of Finnish OCR-treated newspapers from the 19th and early 20th century with a rule-based system (Kettunen et al., 2016) and the Finnish Semantic Tagger, a lexicon-based semantic tagger (Kettunen and Ruokolainen, 2017).

A supervised machine learning system (Nouvel et al., 2011) has been shown to improve F_1–score up to 76.1 percent (Ehrmann et al., 2016). This result was improved furthermore by (Riedl and Padó, 2018) where transfer learning from the German Europeana Newspapers data enabled the biLSTM+CRF classifier to reach a top F_1–score of 78.56 percent (see Table 2).

To conclude, (Schweter and Baiter, 2019) recently employed BERT features for NER resulting in F_1—scores from 75.31 to 79.14 percent while their best models that have been trained on newspaper data of corresponding time epochs deliver F_1—scores from 77.51 to 85.32 percent (see also Table 3).

4 Datasets

4.1 Europeana Newspapers Historic German Datasets

The Europeana Newspapers NER corpus was derived from historical newspapers that have been processed by an OCR and subsequently annotated (Neudecker, 2016). Therefore, that corpus constitutes a good match for the kind of material addressed in this paper. It comprises data sets for historical Dutch, French, and German where the German data has been sourced from newspapers from 1926 from the Dr Friedrich Tessmann Library (LFT), newspapers from 1710 to 1873 from the Austrian National Library (ONB), and newspapers from 1872 to 1930 from the Berlin State Library (SBB).

4.2 CoNLL 2003 German Named Entity Recognition Ground Truth

The German data used in the CoNLL 2003 task (Tjong Kim Sang and De Meulder, 2003) has been taken from a German newspaper, the Frankfurter Rundschau, from 1992. The CoNLL set posesses two different test sets, i.e. TEST-A and TEST-B. We use both in the experiments only for testing (DE-CoNLL-TEST).

4.3 GermEval Konvens 2014 Shared Task Data

The GermEval dataset (Benikova et al., 2014) has been sourced from sampling German Wikipedia and various online newspapers. The GermEval dataset posesses a training, a development and a test set. The development set has not been used at all in the experiments.

4.4 Distribution of Entities

The distribution of labeled entity tokens within the different NER ground truth data sets is shown in Table 1.

	LOC	ORG	PER	Size
DE-CoNLL-TEST	0.025	0.033	0.037	103387
DE-CoNLL-TRAIN	0.025	0.020	0.022	206931
GermEval-TEST	0.028	0.021	0.027	96499
GermEval-TRAIN	0.028	0.022	0.027	452853
LFT	0.062	0.037	0.067	70259
ONB	0.066	0.007	0.115	28012
SBB	0.022	0.010	0.019	47281

Table 1: Distribution of entity tokens amongst different training sets and frequencies of entity tokens across different training sets.

4.5 Digital Collections of the Berlin State Library (DC-SBB)

At the time of the writing of this paper, the digital collections of the SBB contain 153,942 digitized works from the time period of 1470 to 1945 (see Figure 1). Up to now, 28,909 works have been OCR-processed resulting in 4,988,099 fulltext pages.

We applied a sequence of filter steps in order to exclude pages that do not contain german text, have very bad OCR results or contain content that is unlikely to be continuous text.

For each page with OCR text, we predicted its language by means of the *langid* tool (Lui and Baldwin, 2012). Figure 2 illustrates the number of pages per language limited to the most frequent

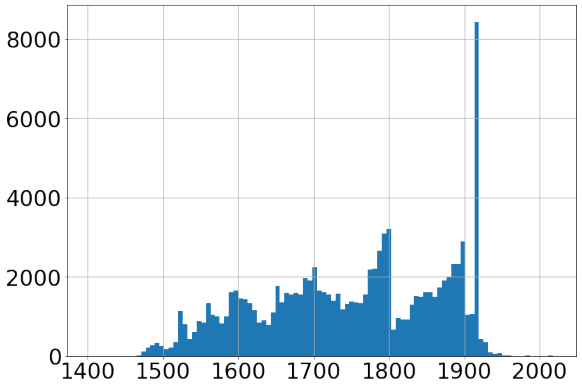

Figure 1: Distribution of publication dates in the digital collections of the Berlin State Library (DC-SBB).

languages. For 19,669 works, the language is consistent over all pages as can be seen from the histogram of detected languages per work that is given in Figure 3. Due to this consistency for the vast majority of all works, we consider the per page language detection provided by langid as sufficiently reliable means to filter out non-german pages. Additionally, we take into account only pages with a confidence score of the German language detection greater than 0.999999.

Fulltexts of pages where the OCR did not work at all, for instance pages that contain hand-written parts, tend to look like random character sequences. In order to exclude these "broken" pages from the data, we computed the distribution of the per-page character entropy rate over all pages. Figure 4 depicts the distribution of the per page character entropy rate in the DC-SBB. We excluded all pages with a character entropy rate below the 0.2 percentile or above the 0.8 percentile of that distribution from the dataset.

As a consequence of these filter steps, 2,333,647 pages of unlabeled historical German text remain and form the DC-SBB dataset. The full dataset is available freely online (Labusch and Zellhöfer, 2019).

5 Experiments

In the scope of the three presented experiments, the BERT model is trained directly with respect to the NER by implementation of the same method that has been proposed by the BERT authors (Devlin et al., 2018). During training, the maximum sequence length is set to 128.

Throughout all experiments, we use the Adam optimizer algorithm with decoupled weight decay (Loshchilov and Hutter, 2019) where the weight

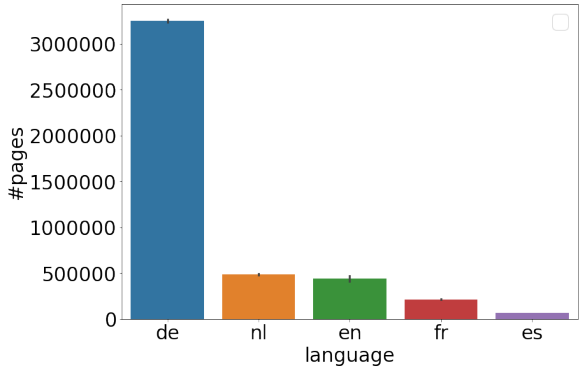

Figure 2: Number of pages per language as detected by *langid* for the most common languages.

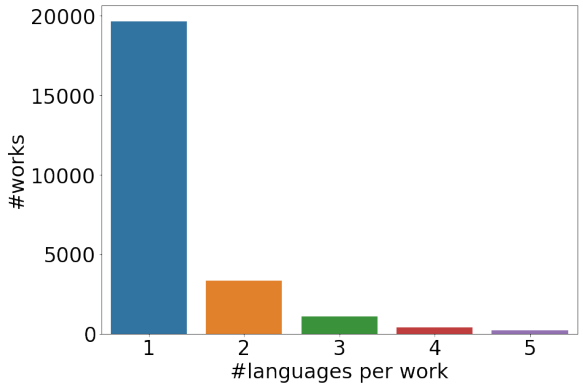

Figure 3: Number of detected languages by *langid* per DC-SBB document (documents with >5 languages are omitted).

decay is set to 0.03. We apply a linear learning rate schedule where warm-up and cool-down of the learning rate take 40% of the performed training steps. We set the target learning rate to $3 * 10^{-5}$ and use a batch size of 32 during all the experiments. We carried out 7 training epochs if not noted otherwise.

Accumulative gradient descent for both supervised and unsupervised learning is applied due to hardware limitations that would otherwise enforce a smaller batch size. Instead of the original BERT implementation, all experimental runs rely on an equivalent PyTorch implementation provided by (Hugging Face, 2019) since accumulative gradient descent cannot be easily carried out using the current Tensorflow (< 2.0) implementation of BERT.

5.1 BERT-Base Multi-Lingual Cased Model

In the first batch of experiments, we explore the NER performance of the baseline model as it has been provided by Google [5]. We use their BERT-Base multi-lingual cased model that has been pre-trained on 104 languages. It has 12 transformer blocks where each transformer block has 768 layers with 12 attention heads and uses a vocabulary size of 119,547. The entire model has about 110 million parameters. The left F_1-column of Table 2 shows the results of the BERT-Base model for different combinations of training and test sets.

[5] https://github.com/google-research/bert/blob/master/multilingual.md

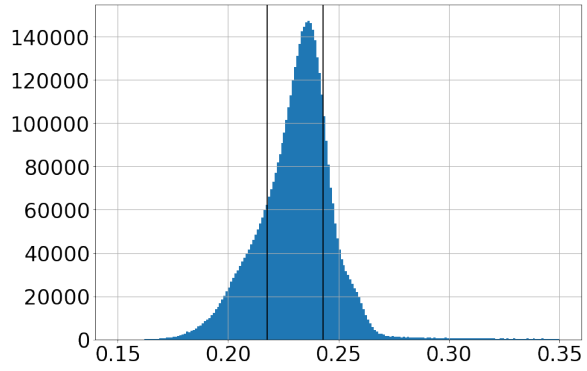

Figure 4: Distribution of the per page character entropy rate of the documents in the DC-SBB dataset. The 0.2 and 0.8 percentiles have been marked with a vertical line.

5.2 BERT-Base Model with Pre-Training on DC-SBB

In this experimental run, we study the impact of unsupervised pre-training with respect to the NER performance on historical and contemporary data. Therefore the multi-lingual BERT-Base model is pre-trained unsupervisedly on the DC-SBB dataset (see Sec. 4.5). The unsupervised pre-training task is composed of the "Masked-LM" and "Next Sentence Prediction" tasks that have been proposed in (Devlin et al., 2018).

The pre-training of the base model has been run for approximately 500 hours on a single NVIDIA 2080 GPU which is equivalent to 5 epochs. During pre-training, the batch size is set to 128, the learning rate is set as in the NER task training and a weight decay of 0.01 is used. The middle F_1-column of Table 2 shows results of the BERT-Base model being pre-trained on the DC-SBB data for different combinations of training and test sets.

5.3 5-fold Cross Validation and Comparison with State of the Art Approaches

Since the NER performance varies heavily for different train/test set combinations and in order to make our results comparable to results in (Riedl and Padó, 2018) and (Schweter and Baiter, 2019), we run a third batch of experiments where a 5-fold

cross validation is performed for the three historical German corpora.

In this run, the impact of pre-training on the model performance under cross validation is evaluated. We apply unsupervised pre-training on the DC-SBB data as well as supervised pre-training on contemporary NER ground truth. In case of supervised pre-training, 7 training epochs are run again with the same learning parameters described above. Finally, unsupervised and supervised pre-training are combined where unsupervised is done first and supervised second. The corresponding cross validation results are shown in Table 3.

6 Discussion

The NER ground truth sets that have been used in the experiments described above are diverse in terms of size and with respect to the frequencies of the entity classes as Table 1 summarizes.

While the contemporary data sets GermEval and CoNLL show similar frequencies of entity classes, the frequencies of entities within the historical data sets LFT, ONB, and SBB deviate significantly. The SBB set comes closest to the contemporary sets in terms of entity frequencies.

Futhermore, there is far more contemporary ground truth available than for historical texts. The amount of ground truth also varies significantly among the various historical datasets.

Table 2 shows the NER performance in terms of the F_1–score obtained with different training/test combinations using either the original BERT-Base model or a BERT-Base model that has been pre-trained on the DC-SBB set. (Riedl and Padó, 2018) present a comprehensive evaluation of CRF and bidirectional long short term memory (biLSTM) with CRF layer approaches for NER in contemporary and historical German, relying on a partial utilization of the ground truth data that is considered in this work. The authors use character embeddings together with different pre-trained word embeddings as input features of the biLSTMs. For those training/test pairs that have corresponding results in (Riedl and Padó, 2018), their best result is listed in the rightmost F_1-column of Table 2.

Interestingly, unsupervised pre-training on DC-SBB data worsens BERT performance in the case of contemporary training/test pairs while the performance improves for all experiments that test on historical ground truth with one exception (CoNLL/LFT). Please note that the same training

		BERT multi-lingual-cased		(Riedl and Padó, 2018)
	pre-train:	none	DC-SBB	none
train	test	F_1	F_1	F_1
CoNLL	CoNLL	**84.5**	82.6	82.99
	LFT	**52.9**	52.0	49.28
	ONB	56.1	56.6	**58.79**
	SBB	67.6	**68.3**	-
GermEval	GermEval	**88.6**	86.7	82.93
	LFT	54.2	54.8	**55.99**
	ONB	60.0	**62.6**	61.35
	SBB	63.1	**65.1**	-
GermEval + CoNLL	CoNLL	**80.2**	79.4	-
	GermEval	**88.0**	85.7	-
	LFT	55.1	**55.2**	-
	ONB	58.6	**60.1**	-
	SBB	64.1	**65.1**	-
LFT	ONB	71.5	**75.9**	65.53
	SBB	54.4	**56.9**	-
LFT+SBB	ONB	72.5	**75.7**	-
ONB	LFT	59.4	**61.5**	49.35
	SBB	51.3	**54.6**	-
ONB+LFT	SBB	54.0	**55.5**	-
ONB+SBB	LFT	61.9	**62.7**	-
SBB	LFT	53.9	**54.9**	-
	ONB	63.4	**66.0**	-

Table 2: BERT NER-performance on different combinations of training and test sets. For all training/test pairs the same number of training epochs has been executed and the same learning parameters have been used.

Left (pre-train none): NER-performance of the non-modified multi-lingual BERT-Base model as provided by Google[5].

Middle (pre-train DC-SBB): NER-performance of the multi-lingual BERT-Base model that has been pre-trained for 5 epochs on the DC-SBB data with objective "Masked-LM" and "Next Sentence Prediction" as proposed in (Devlin et al., 2018) prior to the NER supervised training.

Right (Riedl and Padó, 2018): NER-performance as published in (Riedl and Padó, 2018) where multiple state-of-the art CRF only and biLSTM + CRF approaches using different character and word embeddings have been evaluated.

Pre-training on DC-SBB improves results for historical German datasets, independently on the type of NER-ground-truth used for supervised training whereas the original BERT-base model provides better results on contemporary German test sets.

		BERT multi-lingual-cased			(Riedl and Padó, 2018)	(Schweter and Baiter, 2019)
5-fold cross validation on	pre-train	precision	recall	F_1	F_1	F_1
SBB	DC-SBB + GermEval + CoNLL	81.1 ±1.2	87.8 ±1.4	**84.3 ±1.1**	-	-
	DC-SBB + CoNLL	81.0 ±2.1	87.6 ±1.8	84.2 ±1.9	-	-
	DC-SBB + GermEval	80.6 ±1.8	87.4 ±1.3	83.8 ±1.2	-	-
	CoNLL	81.0 ±1.9	86.6 ±2.2	83.7 ±1.5	-	-
	GermEval	79.7 ±1.8	87.2 ±0.8	83.3 ±1.1	-	-
	GermEval + CoNLL	79.9 ±2.1	86.4 ±1.7	83.0 ±1.9	-	-
	DC-SBB	79.1 ±2.6	86.7 ±0.7	82.7 ±1.3	-	-
	none	79.1 ±3.6	85.0 ±1.1	81.9 ±2.2	-	-
ONB	Newspaper (1703-1875)	-	-	-	-	**85.31**
	DC-SBB+GermEval + CoNLL	81.5 ±1.8	87.8 ±1.4	84.6 ±1.5	-	-
	DC-SBB + GermEval	81.6 ±2.5	87.5 ±1.6	84.5 ±1.8	-	-
	DC-SBB + CoNLL	81.7 ±2.8	87.5 ±1.9	84.5 ±2.3	-	-
	DC-SBB	81.8 ±2.3	87.1 ±2.1	84.3 ±2.0	-	-
	GermEval	80.8 ±2.1	85.4 ±1.2	83.0 ±1.4	78.56	-
	GermEval + CoNLL	80.0 ±1.5	84.7 ±1.6	82.3 ±1.5	-	-
	CoNLL	79.1 ±2.5	84.5 ±2.1	81.7 ±2.2	76.17	-
	none	78.0 ±2.4	84.1 ±1.9	80.9 ±2.0	73.31	-
LFT	Newspaper (1888-1945)	-	-	-	-	**77.51**
	DC-SBB + CoNLL	70.0 ±2.6	81.0 ±0.7	75.1 ±1.5	-	-
	DC-SBB + GermEval	69.9 ±3.0	81.1 ±1.0	75.1 ±1.8	-	-
	DC-SBB	70.0 ±3.5	80.8 ±1.4	75.0 ±2.1	-	-
	DC-SBB + GermEval + CoNLL	69.8 ±3.0	80.8 ±0.9	74.9 ±2.0	-	-
	GermEval	68.9 ±2.7	79.3 ±1.4	73.7 ±1.9	74.33	-
	GermEval + CoNLL	69.1 ±2.6	78.8 ±1.3	73.6 ±1.5	-	-
	none	68.8 ±3.4	79.2 ±1.5	73.6 ±2.2	69.62	-
	CoNLL	68.4 ±3.1	79.1 ±1.3	73.3 ±2.1	72.9	-

Table 3: 5-fold cross validation results for different historical German NER corpora where different pre-training steps have been applied to the BERT model. For all experiments the same number of training epochs and the same learning parameters have been used. Results in (Riedl and Padó, 2018) and (Schweter and Baiter, 2019) have been obtained for some 80/20 training/test split.

None: Model as published by Google[5].

DC-SBB: Model unsupervisedly pre-trained on DC-SBB.

CoNLL: Model supervisedly pre-trained on CoNLL training set.

GermEval: Model supervisedly pre-trained on GermEval training set.

DC-SBB + GermEval + CoNLL: First unsupervised pre-training for 5 epochs on the DC-SBB data. Second supervised pre-training on the joined GermEval and CoNLL NER ground truth.

The NER-performance under cross-validation can be significantly improved by combination of unsupervised and supervised pre-training. DC-SBB+GermEval+CoNLL pre-trained models show close to state-of-the-art performance on all three historical datasets using exactly the same training parameters and number of training epochs.

data leads to significantly different performances on varying test sets.

The original BERT model performs better than the biLSTM+CRF models in the case of contemporary training/test combinations. The pre-trained BERT model performs better than the biLSTM+CRF models in the case of the majority of historical training/test combinations except the CoNLL/ONB and GermEval/LFT pairs.

The impact of the diversity of the ground truth data sets makes it difficult to assess the actual performance of the BERT models on the historical data based on the results shown in Table 2 alone. In order to further study and clarify the experimental outcomes, another sequence of experiments was performed to evaluate the NER performance on the historical data under cross validation. The corresponding results are shown in Table 3. As above, the corresponding best results from (Riedl and Padó, 2018) are listed, if available, though their results have not been obtained under cross validation but for a fixed training/test split. (Schweter and Baiter, 2019) present a recent study of NER in historical German. They use a combined biLSTM + CRF model together with varying combinations of character embeddings, contextualized string embeddings (Akbik et al., 2018), pre-trained word embeddings, and BERT-layer features. We included their best results that have been obtained for a fixed train/test split on the LFT and ONB data set in the rightmost column of Table 3.

As illustrated by Table 3, various degrees of pre-training successively improve the performance of the BERT model. In case of the ONB and LFT data unsupervised pre-training alone (DC-SBB) provides the biggest part of improvement. Additional supervised pre-training adds only a small improvement. In case of the SBB ground truth, which is more similar to the contemporary data, supervised pre-training contributes more to the performance improvement.

BERT outperforms the biLSTM + CRF approaches that have been evaluated in (Riedl and Padó, 2018) but the results are still worse than some of the results reported in (Schweter and Baiter, 2019). Their best results rely on a pre-training scheme that is adapted to the final target domain whereas in our experiments the pre-training scheme DC-SBB + GermEval + CoNLL provides very good cross-validation performance for the three historical German sets SBB, ONB, and LFT while utilizing the same set of learning parameters.

7 Conclusion and Future Work

The historical texts of the SBB digital collections originate from a broad period of time ranging from 1470 to 1945. A long term goal is to reliably conduct NER in this large text corpus in order to improve the user experience for researchers interacting with the library's digitized holdings. Hence, a versatile approach is required that can deliver decent recognition performance for texts of different time epochs and a variety of text categories.

Our results show that an appropriately pre-trained BERT model delivers decent recognition performance in a variety of settings and even provides state of the art performance in many cases without extensive fine-tuning and optimization requirements. This outcome encourages further refinement and an extension of the methodology that has been evaluated in the presented experiments.

In the scope of this paper, we started all our experiments from the BERT-Base model. An increase of the model size is expected to improve the results further (Devlin et al., 2018). Therefore, we plan to re-run the experiments using BERT-Large which requires even more computation time.

In particular, the unsupervised pre-training on the DC-SBB set is computationally demanding. So far, we performed only 5 training epochs though further improvement in the unsupervised tasks "Masked-LM" and "Next Sentence Prediction" is still possible according to the trend of the loss. We plan to compensate for some of the additional computational demand by better and more GPU hardware that is currently installed at the SBB.

We think that there is a lot of performance to gain for historical text by adding more historical ground-truth data. Therefore, we plan to add more historical ground-truth data in the near future also in cooperation with the SoNAR project (Interfaces to Data for Historical Social Network Analysis and Research).

To end with, we plan to significantly reduce the level of noise in the source OCR texts by means of re-processing the digitized documents with LSTM OCR software specifically trained on historical texts and through the application of unsupervised OCR post-correction methods based on neural networks and finite-state-transducers being developed in the OCR-D project (Neudecker et al., 2019).

Acknowledgements

This work was partially supported by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF), project grant QURATOR - Curation Technologies[6].

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

Beatrice Alex and John Burns. 2014. Estimating and rating the quality of optically character recognised text. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, pages 97–102. ACM.

Darina Benikova, Chris Biemann, Max Kisselew, and Sebastian Padó. 2014. Germeval 2014 named entity recognition: Companion paper. *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition, Hildesheim, Germany*, pages 104–112.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT GitHub. `https://github.com/google-research/bert/blob/master/README.md`.

Maud Ehrmann, Giovanni Colavizza, Yannick Rochat, and Frédéric Kaplan. 2016. Diachronic evaluation of ner systems on old newspapers. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 97–107. Bochumer Linguistische Arbeitsberichte.

Maria Federbusch, Christian Polzin, and Thomas Stäcker. 2013. Volltext via OCR- Möglichkeiten und grenzen. *Beiträge aus der Staatsbibliothek zu Berlin - Preußischer Kulturbesitz*, 43:1–138.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL05)*, pages 363–370.

Olivier Galibert, Sophie Rosset, Cyril Grouin, Pierre Zweigenbaum, and Ludovic Quintard. 2012. Extended named entities annotation on OCRed documents: from corpus constitution to evaluation campaign. In *International Conference on Language Resources and Evaluation*, Istanbul, Turkey, January.

[6]https://qurator.ai

Claire Grover, Sharon Givon, Richard Tobin, and Julian Ball. 2008. Named entity recognition for digitised historical texts. *LREC 2008*.

Hugging Face. 2019. BERT PyTorch GitHub. `https://github.com/huggingface/pytorch-pretrained-BERT`.

Kimmo Kettunen and Teemu Ruokolainen. 2017. Names, right or wrong: Named entities in an ocred historical finnish newspaper collection. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, pages 181–186. ACM.

Kimmo Kettunen, Eetu Mäkelä, Teemu Ruokolainen, Juha Kuokkala, and Laura Löfberg. 2016. Old content and modern tools-searching named entities in a finnish ocred historical newspaper collection 1771-1910. *arXiv preprint arXiv:1611.02839*.

Kai Labusch and David Zellhöfer. 2019. OCR Fulltexts of the Digital Collections of the Berlin State Library (DC-SBB), June 26th. `https://doi.org/10.5281/zenodo.3257041`.

Daniel Lopresti. 2009. Optical character recognition errors and their effects on natural language processing. *International Journal on Document Analysis and Recognition (IJDAR)*, 12(3):141–151.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Marco Lui and Timothy Baldwin. 2012. Langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, ACL '12, pages 25–30, Stroudsburg, PA, USA. Association for Computational Linguistics.

Clemens Neudecker, Lotte Wilms, Willem Jan Faber, and Theo van Veen. 2014. Large-scale refinement of digital historic newspapers with named entity recognition. In *Proceedings of the IFLA Newspapers/GENLOC Pre-Conference Satellite Meeting*.

Clemens Neudecker, Konstantin Baierer, Volker Hartmann, Maria Federbusch, Matthias Boenig, and Elisa Hermann. 2019. OCR-D: An end-to-end open source ocr framework for historical printed documents. In *Proceedings of the Third International Conference on Digital Access to Textual Cultural Heritage*, page *in press*. ACM.

Clemens Neudecker. 2016. An open corpus for named entity recognition in historic newspapers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4348–4352, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Damien Nouvel, Jean-Yves Antoine, and Nathalie Friburger. 2011. Pattern mining for named entity recognition. In *Language and Technology Conference*, pages 226–237. Springer.

Thomas L Packer, Joshua F Lutes, Aaron P Stewart, David W Embley, Eric K Ringger, Kevin D Seppi, and Lee S Jensen. 2010. Extracting person names from diverse and noisy ocr text. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, pages 19–26. ACM.

Michael Piotrowski. 2012. Natural language processing for historical texts. *Synthesis lectures on human language technologies*, 5(2):1–157.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *arxiv*.

Martin Riedl and Sebastian Padó. 2018. A named entity recognition shootout for German. In *Proceedings of ACL*, pages 120–125, Melbourne, Australia.

Stefan Schweter and Johannes Baiter. 2019. Towards robust named entity recognition for historic german. *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP)*, page *in press*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 142–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.