An Assessment of the Impact of OCR Noise on Language Models

Konstantin Todorov[a] and Giovanni Colavizza[b]
Institute for Logic, Language and Computation (ILLC), University of Amsterdam, The Netherlands

Keywords: Machine Learning, Language Models, Optical Character Recognition (OCR).

Abstract: Neural language models are the backbone of modern-day natural language processing applications. Their use on textual heritage collections which have undergone Optical Character Recognition (OCR) is therefore also increasing. Nevertheless, our understanding of the impact OCR noise could have on language models is still limited. We perform an assessment of the impact OCR noise has on a variety of language models, using data in Dutch, English, French and German. We find that OCR noise poses a significant obstacle to language modelling, with language models increasingly diverging from their noiseless targets as OCR quality lowers. In the presence of small corpora, simpler models including PPMI and Word2Vec consistently outperform transformer-based models in this respect.

1 INTRODUCTION

Statistical neural language models have become the backbone of modern-day natural language processing (NLP) applications. They have proven high capabilities in learning complex linguistic features and for transferable, multi-purpose adaptability Qiu et al. (2020), in particular with the recent success of contextual models like BERT Devlin et al. (2019). Language models' main objective is to assign probabilities to sequences of linguistic units, such as sentences made of words, possibly benefiting from auxiliary tasks. The success of neural language models has fostered a significant amount of work on understanding how they work internally Rogers et al. (2020). In Digital/Computational Humanities, language models are primarily used as components of NLP architectures and to perform well-posed tasks. Examples include Handwritten/Optical Character Recognition (H/OCR) Kahle et al. (2017), Named Entity Recognition (NER) and linkage Ehrmann et al. (2020), modelling semantic change Shoemark et al. (2019), annotating historical corpora Coll Ardanuy et al. (2020), translating heritage metadata Banar et al. (2020), and many more.

An open challenge for neural language models are low-resource settings: languages or tasks where language data is comparatively scarce, and where annotations are few Hedderich et al. (2021). This is a known issue for many underrepresented languages, including when a language is distinctively appropriated for example via dialects and idiomatic expressions Nguyen et al. (2016). Consequences include the possible exclusion of speakers of low-resource languages, cognitive and societal biases, the reduction of linguistic diversity and often poor generalization Ruder (2020). Historical language data are also comparatively less abundant and sparser than modern-day data Piotrowski (2012); Ehrmann et al. (2016).

What is more, historical language data often poses two additional challenges: noise and variation. Noise comes from errors which should be corrected and their impact mitigated, variation instead is a characteristic of language which may constitute a useful signal. Modern-day NLP methods, including language models, 'overcome' noise by sheer data size – yet noise can still remain a problem – and often flatten-out linguistic variation. Therefore, when employed on real-world applications in low-resource settings, or when applied out-of-domain, these methods might fall short and, crucially, they might fail to appropriately deal with noise and variation.

In this work, we contribute to bridge this gap by posing the following question: *what is the impact on language models of textual noise caused by OCR?*

While recent work has focused on the impact of OCR noise on downstream tasks Hill and Hengchen (2019); van Strien et al. (2020); Todorov and Colavizza (2020), less is known about language mod-

[a] https://orcid.org/0000-0002-7445-4676
[b] https://orcid.org/0000-0002-9806-084X

Todorov, K. and Colavizza, G.
An Assessment of the Impact of OCR Noise on Language Models.
DOI: 10.5220/0010945100003116
In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022) - Volume 2*, pages 674-683
ISBN: 978-989-758-547-0; ISSN: 2184-433X

els in this respect. We therefore proceed by considering a most basic empirical setting, expanding from van Strien et al. (2020). First, we use data in multiple languages available in two versions: as ground truth, assumed correct and verified by humans, and as OCRed texts, containing varying degrees of noise from automated text extraction. The data we use is small when compared to modern-day language modelling standards, yet of realistic size in a Digital/Computational Humanities setting. Furthermore, we consider language models trained from scratch and do not cover fine-tuning or language model adaptation here. Lastly, for each language model under consideration we compare the results of two identically-configured language models, trained on these two versions of the same corpus, by inspecting how similar their learned vector spaces are upon convergence. In this way, we assume no universal baseline, but instead compare language models independently.

While transfer learning on language models is of extreme importance to model NLP applications Ruder (2019), we do not consider it here. The reason is the difficulty in establishing a consistent comparison. While comparing the vector spaces of language models trained from scratch is possible upon convergence, this is more problematic for fine tuning since it is difficult to know when a fine tuned model has actually converged to an accurate model of the new domain. This issue is particularly severe when using small datasets to fine tune. In fact, in the evaluation setting described above, the best result would be achieved by performing no fine tuning at all. Indeed, in this case, both the ground truth and the OCR models would be perfectly identical. A way around is to use extrinsic evaluation, and consider (the similarity in performance on) downstream tasks instead. While extrinsic evaluation has its merits, it does not allow to perform a direct assessment of a language model, but only one limited to its usefulness for downstream tasks. We therefore consider extrinsic evaluation to be complementary to the approach we pursue here.

2 RELATED WORK

Neural Language Models. Vector representations of linguistic units, referred to as embeddings, have been instrumental in the success of modern neural NLP. On the one hand, as statistical models of language when trained on unsupervised objectives and, on the other hand, as components in larger NLP architectures Xia et al. (2020); Qiu et al. (2020). Very popular models include Word2Vec Mikolov et al. (2013b,a) and BERT Devlin et al. (2019). A common theme of neural language modelling research over time seems to be that increasing larger parameters and datasets lead to better results Brown et al. (2020); Raffel et al. (2020). More recently, attention is increasing for low-resource languages which do not yet possess the amount of data or resources which are readily available for, say, English Ruder (2020); Hedderich et al. (2021). As a consequence, promising work is emerging on effective small language models Schick and Schütze (2021).

Language Models and Noise. A challenge in language modelling which is often – yet not necessarily – occurring in low-resource settings is noise. We can consider noise as unwanted errors in the texts, introduced by processing steps. Examples include errors in audio to text recognition or in transcription. Noise can be caused by humans, machines, communication channels; it can be systematic or not. Substantial work on noise and language models has so far focused on robustness to adversarial attacks Pruthi et al. (2019). Some approaches to language modelling can indeed be more resilient to noise than others, despite often having been introduced for other reasons (primarily dealing with out-of-vocabulary words and being language agnostic). BERT, for example, has been proven sensitive to (non-adversarial) human noise Sun et al. (2020); Kumar et al. (2020). Examples of models that can be more resilient to noise include typological language models Gerz et al. (2018); Ponti et al. (2019), sub-word or character-level language models Kim et al. (2016); Zhu et al. (2019); Ma et al. (2020), byte-pair encoding Sennrich et al. (2016), and their extension in recent tokenization-free models (Heinzerling and Strube, 2018; Clark et al., 2021; Xue et al., 2021), yet their use as noise-resilient language models remains to be fully assessed.

Assessing the Impact of OCR Noise. A growing body of work is focused on assessing and mitigating the impact of OCR noise. An area of active work is that of post-OCR correction, which attempts to automatically improve on a noisy text after OCR Hämäläinen and Hengchen (2019); Boros et al. (2020); Nguyen et al. (2020); Todorov and Colavizza (2020). Several recent contributions have assessed the impact of OCR noise using extrinsic evaluation and considering a variety of downstream tasks, including topic modelling, text classification, named entity recognition and information retrieval, among others Hamdi et al. (2019); Hill and Hengchen (2019); van Strien et al. (2020); Boros et al. (2020). While more systematic comparisons are needed, the general

trend seems to indicate that OCR texts are often 'good enough' to be used for downstream tasks.

It is our intent to contribute to this growing body of work by assessing the impact of OCR noise on a selection of mainstream language models, in view of informing their use and future development.

3 EXPERIMENTAL SETUP

In this section we introduce the language models which we consider for this study and present the data we used for our experiments. Lastly, we detail the evaluation procedure to assess their resilience to OCR noise.

3.1 Language Models

There are several popular techniques for language modelling. Modern-day language models are typically vector-based: they map linguistic units (e.g., tokens) to vectors of a given size. How these vectors are updated and learned during training depends on the model at hand and the task(s) it focuses on. In order to compare different and popular approaches, in this study we consider Word2Vec, namely Skip-Gram with Negative Sampling (SGNS) and Continuous Bag-of-words (CBOW) Mikolov et al. (2013b,a), and GloVe Pennington et al. (2014). Furthermore, we consider attention-based models able to make use of the context of an occurrence at inference time and not just at training time, namely BERT Devlin et al. (2019) and ALBERT Lan et al. (2020) (the latter a very fast variant of the former). Finally, we also include a language model based on co-occurrence counts re-weighted using Positive Pointwise Mutual Information (PPMI), as a baseline Church and Hanks (1989).

PMI. is a measure of association that quantifies the likelihood of a co-occurrence of two words taking into account the independent frequency of the two words in the corpus. Positive PMI (PPMI) leaves out negative values, under the assumption that they might capture unreliable and therefore uninformative statistics about the word pair. Formally, PMI and PPMI are calculated as follows:

$$\mathrm{PMI}(w,c) = \log_2 \frac{P(w,c)}{P(w)P(c)} \quad (1)$$

$$\mathrm{PPMI}(w,c) = \max\left(\mathrm{PMI}(w,c), 0\right) \quad (2)$$

Where $P(w,c)$ is the probability of the word w occurring together with word c, and the denominator contains the individual probabilities of these words occurring independently.

Skip-gram. positions each word in a vector space such that similar words are closer to each other. This is achieved by using the self-supervised task of predicting the *context words* given a specific *target word*. To improve the speed of convergence and ease computation, we use *negative sampling*. If we consider the original target and context words as positive examples, we sample negative ones at random from the corpus during training.

CBOW. uses the mirror task compared to Skip-gram, by predicting the target word using context words. For both models, the amount of context words is configurable and called *window size* of the model.

GloVe. is similar to PPMI, in that training is performed on the aggregated word-to-word co-occurrence matrix from a corpus, following the intuition that these encode a form of meaning.

BERT. stands for Bidirectional Encoder Representations for Transformers that, at the time of introduction, gave state-of-the-art results on a wide variety of tasks. BERT is a multi-layered bi-directional transformer encoder. It uses self-supervised tasks such as masked language modelling and next sentence prediction. We apply the former to train our BERT model. The original model uses sub-word tokenization and it generates contextualised word representation at inference time.

ALBERT. uses most of BERT's design choices. However, this model differs by introducing two parameter-reduction techniques that lower memory consumption and drastically decrease training time.

We implement the overall pipeline for our experiments and *PPMI* ourselves. For *CBOW* and *SGNS* we use a popular Python library, *Gensim* Rehurek and Sojka (2010), while for *BERT* and *ALBERT* we rely on *HuggingFace* Wolf et al. (2020). We remark again that we always do training from scratch, without using any pre-trained model.

For all models, we use standard default for the used architecture tokenization. For Skip-gram, CBOW and GloVe we use simple rules, such as cleaning digits, punctuation and multiple white-spaces. For BERT and ALBERT we use Byte-level BPE tokenizer, as introduced by OpenAI and which works on sub-word level (Wang et al., 2020). For these models,

we additionally split the data at 500 characters due to the design limitations in the original implementation. Finally, we use default configurations for the transformers, namely a *vocabulary size* of 30,522 for BERT and of 30,000 for ALBERT, and an *vector size* equal to 768. For ALBERT, we set the number of hidden layers and attention heads to 12 and the intermediate size to 3072 so that these are equal to their counterparts in BERT.

3.2 Data

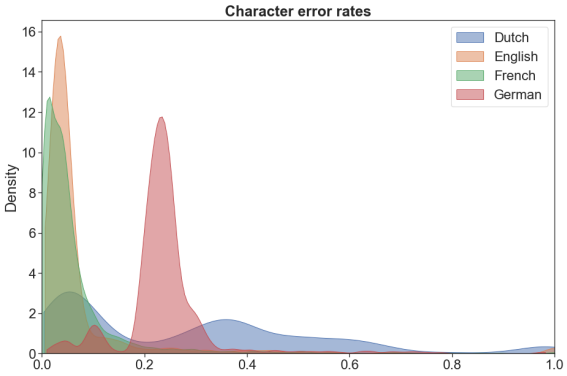
We make use of the datasets provided by the International Conference on Document Analysis and Recognition (ICDAR) 2017 Chiron et al. (2017) and 2019 Rigaud et al. (2019) competitions on Post-OCR text correction. These two datasets combined include ten European languages of which we consider four in this study. The data is provided in three versions: OCR, aligned OCR and ground-truth. We make use of the aligned OCR and ground-truth versions for the purpose of this study and combine the training and evaluation sets together.

From Table 1 we show how different the four languages are in terms of corpus size. Dutch and English languages contain comparatively fewer documents but of longer average size, while French and German have more, usually shorter documents. In order to assess whether having more data for languages with a smaller corpus would alter our results, we experimented with adding data from the National Library of Australia's Trove newspaper archive [1] to the English corpus, but discovered that it does not lead to any differences in the outcomes of our experiments when tested on several of our configurations. We therefore leave this out and only report results using ICDAR 2017+2019 data in what follows.

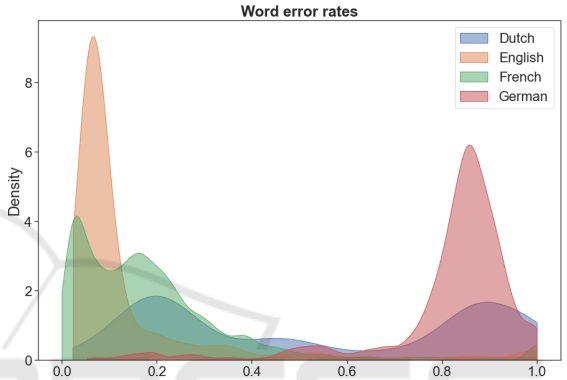
OCR error rates, per language and averaged over documents are given in Table 2, alongside the distribution of character error rates in Figure 1a and of word error rates in Figure 1b. The error rates on character level are calculated by comparing characters on the same position in the OCR and aligned ground-truth versions. Word error rates are calculated following the *de facto* standard approach of word errors to processed words Morris et al. (2004). Documents which are having misaligned OCR and ground-truth versions are excluded from the error rates calculation. It is worth noting that these represent a significant proportion for the German language (Table 1).

Before using each corpus for training, we perform the following pre-processing steps for all of our configurations *except for BERT and ALBERT*. We remove

[1] https://trove.nla.gov.au.



(a) Character-level



(b) Word-level

Figure 1: OCR error rates per language, as distribution of document scores.

numbers and punctuation from the data, lowercase all characters, substitute multiple white-spaces with one, also removing leading and trailing white-spaces in the process, and finally split the different words into tokens. We then build the vocabulary for each version of each corpus – ending up with two versions per corpus: OCR and ground truth – and remove all tokens that occur less than five times overall (per version). We pick this number following previous research and in order to reduce the computational requirements of training our models. For transformer-based models instead, we only split words into sub-word tokens and replace multiple white-spaces with one. These are typical pre-processing steps in view of presenting as much contextual information as possible to the model.

3.3 Evaluation

Our goal is to compare two versions of the same model configuration, each trained from scratch, on the OCR and the ground-truth versions of the same corpus. We remind the reader that our evaluation therefore does not provide any indication of the relative

Table 1: Dataset statistics calculated on ground truth corpora versions. The column *Aligned* reports the number of documents where OCR versions and ground truth are perfectly aligned. The column *Split* reports the number of documents after document splitting for transformer-based models, which require equally-sized data points.

	Documents			**Characters per doc.**			**Total**
	Total	Aligned (% of total)	Split	Avg	Min	Max	**characters**
Dutch	150	149 (99.3%)	5346	4593	42	16,028	688,934
English	963	951 (98.8%)	51,689	6866	2	869,953	6,612,108
French	3993	3616 (90.6%)	84,676	2660	2	195,177	10,620,966
German	10,032	1738 (17.3%)	128,662	1581	126	16,187	15,856,445

Table 2: OCR error rates per language, averaged over documents.

Language	**Error rate level**	
	Character	**Word**
Dutch	0.286	0.536
English	0.075	0.146
French	0.064	0.193
German	0.240	0.813

benefit of using one language model versus another in terms of their capacity to represent language or perform on downstream tasks. Rather, we assess and compare the resilience of each model to OCR noise.

Our evaluation procedure is composed of the following steps:

1. For each corpus/language, we start by taking the intersection of the words that are part of all vocabularies/models. For transformer-based models, which do not have word-level vocabularies, we use the vocabulary intersection from the other models.
2. For transformer-based models, we output the hidden states for all words from the intersected vocabulary. For words which are split into multiple sub-word tokens due to BERT and ALBERT tokenization, we take the mean values. This approach leaves out the contextual information from the inferred embeddings, yet it ensures proper comparison with different architectures.
3. For each corpus/language, model configuration, and word in the vocabulary intersection, we compute the cosine similarity with every other word in the vocabulary intersection.
4. We then compare the amount of overlap in the top N fraction of closest words (neighbors) in the two versions of the same corpus (OCR and ground truth). In this way, we are able to assess to what extent the two models agree on what is the neighborhood for each word in the vocabulary intersection.
5. Since different corpora/languages possess varying vocabulary sizes, we use the percentage over the total dataset-specific vocabulary intersection. N is thus ranging from 0.01 (1%) to 1.0 (100% top neighbors).
6. Following Gonen et al. (2020), we use neighbor overlap as our main evaluation metric. That is the proportion of overlapping neighbours for any word in the vocabulary intersection, defined using the following formula:

$$\text{overlap@}k(r_{OCR}, r_{truth}) = \frac{|r^k_{OCR} \cap r^k_{truth}|}{k} \quad (3)$$

   Where r^k_{OCR} are the top-k neighbors of word r in the *OCR* model, and r^k_{truth} in the ground *truth* model respectively. k is taken to correspond to the top N fraction of neighbors for each specific vocabulary intersection.
7. For example, if we have a vocabulary intersection of size 1000 and evaluate at $N = 0.01$, we would take the $k = 10$ closest words for each word. If on average 5 out of 10 words correspond in the two versions/models, we would have an average 0.5 overlap score (Equation 3).

Further empirical settings are as follows. In agreement and to ensure compatibility with previous work on Word2Vec embeddings, we use embedding size of 300 for English, German and French languages and 320 for Dutch Tulkens et al. (2016). We use the AdamW optimizer Loshchilov and Hutter (2018) for transformer based models. We further assess models using two learning rates – which we label *fast* and *slow* – that are comparatively higher/lower in order to verify the effect of learning speed on our evaluation. Since BERT and ALBERT usually benefit from lower learning rates compared to simpler models such as CBOW and Skip-gram, we adjust accordingly. The learning rates which correspond to fast and slow for each model are shown in Table 3, they have been selected following commonly used default values in the literature.

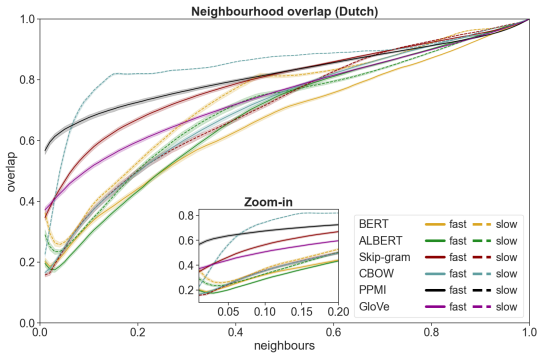


Figure 2: Neighbourhood overlaps (Dutch). 95% bootstrapped confidence intervals are provided at .01 intervals.

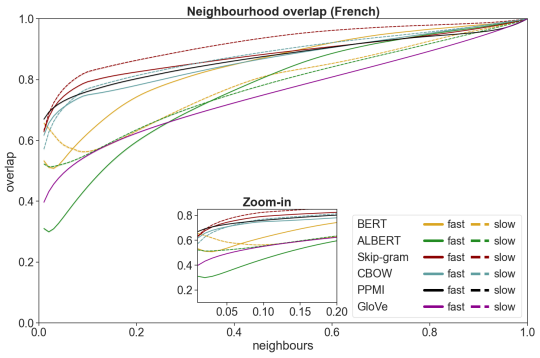


Figure 4: Neighbourhood overlaps (French). 95% bootstrapped confidence intervals are provided at .01 intervals.

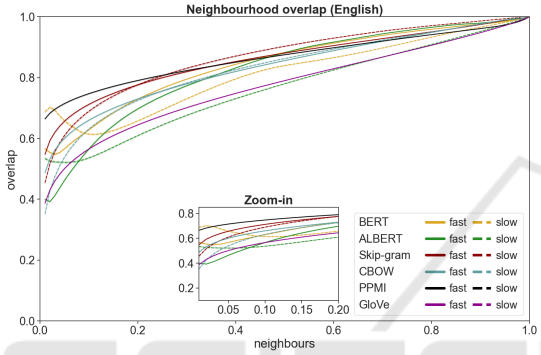


Figure 3: Neighbourhood overlaps (English). 95% bootstrapped confidence intervals are provided at .01 intervals.

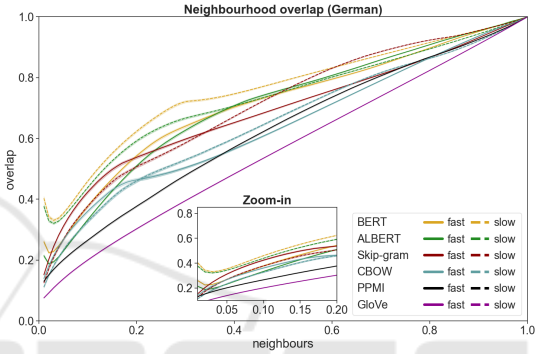


Figure 5: Neighbourhood overlaps (German). 95% bootstrapped confidence intervals are provided at .01 intervals.

To control for the stochasticity of the training process, we show all results by averaging three different runs for each hyper-parameter configuration and model type.

Table 3: Model parameter size comparison and default learning rates.

Model	Number of parameters	Learning rates Fast	Slow
PPMI	N/A		
GloVe	N/A		
CBOW	1.7M	1e−3	1e−4
Skip-Gram	7.5M		
BERT	108M	1e−4	1e−5
ALBERT	12M		

4 RESULTS

Our results are shown in a series of mirrored plots, one per language. The plots show the neighborhood overlap (Eq. 3) on the y axis, at varying values of N from 0.01 (1%) to 1.0 (100%), on the x axis. Naturally, higher values of N are somewhat less interesting and more trivial than lower values of N (overlap of the most similar neighbors), therefore each plot also shows a zoom-in inset for values on N between 0.01 (1%) and 0.2 (20%). Confidence intervals are added and point to the significance of all results.

Starting with Dutch, in Figure 2, we can appreciate how CBOW (slow learning rate) and PPMI appear to be most resilient to OCR noise, followed by Skip-gram (fast learning rate) and GloVe. Every other model configuration performs much worse at lower values of N. Importantly, the Dutch corpus we use is small in comparison to other corpora, and contains poor quality OCR.

Next, we show results for English in Figure 3. The English corpus has a good quality OCR and is of average size in our study. In this setting, PPMI followed by Skip-gram models perform best, notwithstanding a good performance of BERT (slow learning rate) at very low values of N. All other models play catch-up.

The results are quite clear for French, in Figure 4, which is a corpus of size and quality comparable to those of English. For French, Skip-gram, CBOW and PPMI models perform consistently better in resisting the impact of OCR noise. Compared, contextual

BERT and ALBERT are lagging behind, along with GloVe.

Lastly, we report results for German in Figure 5. The German corpus is the largest in size, but the worst in terms of OCR quality. The low OCR quality clearly shows in the comparatively lower overlap scores overall. Surprisingly, for German the best performing models are BERT and ALBERT, in particular on a slow learning rate. It is worth noting that a fast Skip-gram configuration is also close.

Our results show clear trends. Firstly, the impact of OCR quality is overall quite significant: all models trained on OCR data diverge from ground truth. Language models trained on lower OCR quality corpora consistently reach lower overlap scores, as shown for Dutch and German. For example, the best performing models for French, which has good OCR quality, reach over 80% overlap at low N (5%), while this overlap score is only reached at very high values for German (55%). The effect of the size of a corpus does not seem to be significant, yet results for German, the only language where transformer-based models outperform the competition, warrant further scrutiny in this respect. In terms of language models, simpler seems to be better. PPMI, Skip-gram and CBOW models consistently perform above transformer-based models (BERT, ALBERT) in terms of resilience to OCR noise. The only exception in this respect is German, which is plagued by OCR errors but is also the largest corpus available. Furthermore, BERT consistently outperforms ALBERT, hinting at the fact that the gains in training speed come at a cost. GloVe does not appear to be particularly resilient to OCR noise either. Lastly, the variety and lack of consistency of results when using fast and slow learning rates underlines the importance of choosing the right hyper-parameters.

We end our results by highlighting a set of limitations which constitute interesting directions for future work. Firstly, our results hold within the remits of the datasets and models we focused on: using more aligned data – which is unfortunately costly to create –, in more languages of comparable corpus sizes, and with more varied contents and OCR quality would all be useful future contributions. Furthermore, the quality of the ground truth itself, which we took here for granted, might warrant further scrutiny. Future work could also focus on hyper-parameter fine tuning in order to reach the maximum performance with any given model, as well as on assessing how many data are necessary to reach a certain desired result with a given language model. Next, as we discussed at the beginning, our results would benefit from a complementary study focused on the extrinsic evaluation of the impact of OCR on language models, in particular when used as components for machine learning architectures focused on downstream tasks. Related to this point, we decided not to consider pre-trained models: approaching the same research question via extrinsic evaluation would allow to overcome such limitation.

5 CONCLUSION

We have assessed the impact of OCR noise on a variety of language models. Using data in Dutch, English, French and German, of different sizes and OCR qualities, we considered two aligned versions of the same corpus, one OCRed and one manually corrected (ground truth). Two identical instances of each language model were in turn trained from scratch over these two versions of the same corpus, and the similarity of the resulting vector spaces was assessed using word neighborhood overlap. This approach allowed us to assess and compare the resilience of each language model to OCR noise, independently.

We show that OCR noise significantly impacts language models, with a steep degradation of results for corpora with lower OCR quality. Furthermore, we show that 'simpler' language models, including PPMI and the Word2Vec family (Skip-gram and CBOW) are often more resilient to OCR noise than recent transformer-based models (BERT, ALBERT). We also show that the choice of key hyper-parameters, such as the learning rate, significantly impacts results as well. The size of a corpus might also be an important factor, as suggested by transformer-based models performing best with the largest corpus available (German), yet more experiments are needed to untangle its effects.

While several limitations and opportunities for future work have been discussed, we believe the two most important next steps to be the development of multilingual evaluation corpora of similar size and OCR quality in order to study the impact of OCR noise more systematically, and the need to complement our study with an extrinsic assessment on downstream tasks making use of (possibly pre-trained) language models. Nevertheless, we have shown that OCR noise poses a significant obstacle to language modelling. In the presence of small datasets, simpler models including PPMI and Word2Vec are more resilient to OCR noise than transformer-based models trained from scratch.

CODE AND DATA AVAILABILITY

All data is openly provided by the International Conference on Document Analysis and Recognition (ICDAR) 2017 Chiron et al. (2017) and 2019 Rigaud et al. (2019) competitions on Post-OCR text correction.

Our code base is publicly available and described at https://doi.org/10.5281/zenodo.5799211 (Todorov and Colavizza, 2021).

REFERENCES

Banar, N., Lasaracina, K., Daelemans, W., and Kestemont, M. (2020). Transfer Learning for Digital Heritage Collections: Comparing Neural Machine Translation at the Subword-level and Character-level. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, pages 522–529, Valletta, Malta. SCITEPRESS - Science and Technology Publications.

Boros, E., Hamdi, A., Linhares Pontes, E., Cabrera-Diego, L. A., Moreno, J. G., Sidere, N., and Doucet, A. (2020). Alleviating Digitization Errors in Named Entity Recognition for Historical Documents. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 431–441, Online. Association for Computational Linguistics.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Chiron, G., Doucet, A., Coustaty, M., and Moreux, J.-P. (2017). ICDAR 2017 Competition on Post-OCR Text Correction. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1423–1428. IEEE.

Church, K. and Hanks, P. (1989). Word Association Norms, Mutual Information, and Lexicography. In *27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Clark, J. H., Garrette, D., Turc, I., and Wieting, J. (2021). Canine: Pre-training an Efficient Tokenization-Free Encoder for Language Representation.

Coll Ardanuy, M., Nanni, F., Beelen, K., Hosseini, K., Ahnert, R., Lawrence, J., McDonough, K., Tolfo, G., Wilson, D. C., and McGillivray, B. (2020). Living Machines: A study of atypical animacy. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4534–4545, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ehrmann, M., Colavizza, G., Rochat, Y., and Kaplan, F. (2016). Diachronic Evaluation of NER Systems on Old Newspapers.

Ehrmann, M., Romanello, M., Flückiger, A., and Clematide, S. (2020). Overview of CLEF HIPE 2020: Named Entity Recognition and Linking on Historical Newspapers. In Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., Cappellato, L., and Ferro, N., editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, volume 12260, pages 288–310. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Gerz, D., Vulić, I., Ponti, E. M., Reichart, R., and Korhonen, A. (2018). On the Relation between Linguistic Typology and (Limitations of) Multilingual Language Modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327, Brussels, Belgium. Association for Computational Linguistics.

Gonen, H., Jawahar, G., Seddah, D., and Goldberg, Y. (2020). Simple, Interpretable and Stable Method for Detecting Words with Usage Change across Corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 538–555. Association for Computational Linguistics.

Hämäläinen, M. and Hengchen, S. (2019). From the Paft to the Fiiture: a Fully Automatic NMT and Word Embeddings Method for OCR Post-Correction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 431–436, Varna, Bulgaria. INCOMA Ltd.

Hamdi, A., Jean-Caurant, A., Sidere, N., Coustaty, M., and Doucet, A. (2019). An Analysis of the Performance of Named Entity Recognition over OCRed Documents. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 333–334, Champaign, IL, USA. IEEE.

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., and Klakow, D. (2021). A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. *arXiv:2010.12309 [cs]*. arXiv: 2010.12309.

Heinzerling, B. and Strube, M. (2018). BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Hill, M. J. and Hengchen, S. (2019). Quantifying the impact of dirty OCR on historical text analysis: Eighteenth

Century Collections Online as a case study. *Digital Scholarship in the Humanities*, 34(4):825–843.

Kahle, P., Colutto, S., Hackl, G., and Muhlberger, G. (2017). Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pages 19–24, Kyoto. IEEE.

Kim, Y., Jernite, Y., Sontag, D., and Rush, A. M. (2016). Character-Aware Neural Language Models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2741–2749. AAAI Press. event-place: Phoenix, Arizona.

Kumar, A., Makhija, P., and Gupta, A. (2020). Noisy Text Data: Achilles' Heel of BERT. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 16–21, Online. Association for Computational Linguistics.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv:1909.11942 [cs]*. arXiv: 1909.11942.

Loshchilov, I. and Hutter, F. (2018). Fixing Weight Decay Regularization in Adam.

Ma, W., Cui, Y., Si, C., Liu, T., Wang, S., and Hu, G. (2020). CharBERT: Character-aware Pre-trained Language Model. *Proceedings of the 28th International Conference on Computational Linguistics*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In Bengio, Y. and LeCun, Y., editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Morris, A. C., Maier, V., and Green, P. (2004). From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *INTERSPEECH*, page 4.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., and de Jong, F. (2016). Computational Sociolinguistics: A Survey. *Computational Linguistics*, 42(3):537–593.

Nguyen, T. T. H., Jatowt, A., Nguyen, N.-V., Coustaty, M., and Doucet, A. (2020). Neural Machine Translation with BERT for Post-OCR Error Detection and Correction. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 333–336, Virtual Event China. ACM.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, volume 14, pages 1532–1543.

Piotrowski, M. (2012). Natural Language Processing for Historical Texts. *Synthesis Lectures on Human Language Technologies*, 5(2):1–157.

Ponti, E. M., O'Horan, H., Berzak, Y., Vulić, I., Reichart, R., Poibeau, T., Shutova, E., and Korhonen, A. (2019). Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. *Computational Linguistics*, 45(3):559–601.

Pruthi, D., Dhingra, B., and Lipton, Z. C. (2019). Combating adversarial misspellings with robust word recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5582–5591, Florence, Italy. Association for Computational Linguistics.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., and Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Rehurek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks*, pages 46–50, Valletta, Malta. University of Malta.

Rigaud, C., Doucet, A., Coustaty, M., and Moreux, J.-P. (2019). ICDAR 2019 Competition on Post-OCR Text Correction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1588–1593. IEEE.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Ruder, S. (2019). *Neural Transfer Learning for Natural Language Processing*. PhD thesis, National University of Ireland, Galway.

Ruder, S. (2020). Why You Should Do NLP Beyond English. http://ruder.io/nlp-beyond-english.

Schick, T. and Schütze, H. (2021). It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Shoemark, P., Liza, F. F., Nguyen, D., Hale, S., and McGillivray, B. (2019). Room to Glo: A Systematic Comparison of Semantic Change Detection Approaches with Word Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 66–76, Hong Kong, China. Association for Computational Linguistics.

Sun, L., Hashimoto, K., Yin, W., Asai, A., Li, J., Yu, P., and Xiong, C. (2020). Adv-BERT: BERT is not robust on misspellings! Generating nature adversarial samples on BERT.

Todorov, K. and Colavizza, G. (2020). Transfer Learning for Historical Corpora: An Assessment on Post-OCR Correction and Named Entity Recognition. In *Proceedings of the Workshop on Computational Humanities Research (CHR 2020) : Amsterdam, the Netherlands, November 18-20, 2020*, CEUR Workshop Proceedings, 1613-0073, 2723, pages 310–339, Amsterdam. Aachen: CEUR-WS.

Todorov, K. and Colavizza, G. (2021). Zenodo, ktodorov/historical-ocr, An assessment of the impact of OCR noise on language models.

Tulkens, S., Emmery, C., and Daelemans, W. (2016). Evaluating Unsupervised Dutch Word Embeddings as a Linguistic Resource. In Chair), N. C. C., Choukri, K., Declerck, T., Grobelnik, M., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

van Strien, D., Beelen, K., Ardanuy, M., Hosseini, K., McGillivray, B., and Colavizza, G. (2020). Assessing the Impact of OCR Quality on Downstream NLP Tasks. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, pages 484–496, Valletta, Malta. SCITEPRESS - Science and Technology Publications.

Wang, C., Cho, K., and Gu, J. (2020). Neural machine translation with byte-level subwords. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9154–9160.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and et al. (2020). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771 [cs]*. arXiv: 1910.03771.

Xia, P., Wu, S., and Van Durme, B. (2020). Which *BERT? A Survey Organizing Contextualized Encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7516–7533, Online. Association for Computational Linguistics.

Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., and Raffel, C. (2021). Byt5: Towards a token-free future with pre-trained byte-to-byte models.

Zhu, Y., Heinzerling, B., Vulić, I., Strube, M., Reichart, R., and Korhonen, A. (2019). On the Importance of Subword Information for Morphological Tasks in Truly Low-Resource Languages. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 216–226, Hong Kong, China. Association for Computational Linguistics.