Assessing and Minimizing the Impact of OCR Quality on Named Entity Recognition

Ahmed Hamdi, Axel Jean-Caurant, Nicolas Sidere,
Mickaël Coustaty, Antoine Doucet

University of La Rochelle, L3i Laboratory, France
{firstname.lastname}@univ-lr.fr

Abstract. In digital libraries, the accessibility of digitized documents is directly related to the way they are indexed. Named entities are one of the main entry points used to search and retrieve digital documents. However, most digitized documents are indexed through their OCRed version and OCR errors may hinder their accessibility. This paper aims to quantitatively estimate the impact of OCR quality on the performance of named entity recognition (NER). We tested state-of-the-art NER techniques over several evaluation benchmarks, and experimented with various levels and types of synthesised OCR noise so as to estimate the impact of OCR noise on NER performance. We share all corresponding datasets. To the best of our knowledge, no other research work has systematically studied the impact of OCR on named entity recognition over datasets in multiple languages. The final outcome of this study is an evaluation over historical newspaper data of the national library of Finland, resulting in an increase of around 11 percentage points in terms of F1-measure over the best-known results to this day.

Keywords: Digitized documents · Indexing · OCR · Named entity recognition.

1 Introduction

Substantial amounts of printed documents are digitized and archived as images in digital libraries. This is notably the case of historical documents, which require an *Optical Character Recognition* (OCR) step to give access to their textual content. Unfortunately, while the performance of OCR systems has greatly improved, it remains imperfect. In addition, a great deal of documents were digitized in a time when storing high-quality images was difficult. Such documents cannot readily benefit from improvements in OCR quality. Several studies understandably suggest that the performance of natural language processing tools is harmed by the use of OCRed text, i.e., text resulting from an OCR process [18]. This naturally makes document access more difficult since simple keyword search will for instance not match a query with the corresponding words if they suffer from OCR errors. The quality of the text generated using OCR engines depends on the algorithms used in OCR, on the parameter settings of the scanner used to digitize documents, on the quality of the original image and on the

nature of the document. For instance, text generated from recent vs. historical newspapers or well-preserved vs. damaged manuscripts is usually not of the same quality. Even though a reasonable amount of OCR errors is known to have low impact on the readability of documents, the errors will be indexed as they are by search engines and other NLP tools. Subsequently, if some words are incorrectly recognized by the OCR process, they will be indexed with their errors. This causes a chain reaction for tools developed to analyze the resulting content.

A study has shown that named entities (NEs) are the first point of entry for users in a search system [10]. As an illustration, it has been observed that 4 out of 5 user queries on the Gallica digital library [1] contain at least one named entity [2]. For this reason, their quality is far more critical than that of most other words in OCRed documents. In order to improve the satisfaction of users' information needs, it is thus necessary to ensure their quality.

Named entity recognition (NER) is a task that emerged in the middle of the 1990*s* [12]. It aims to locate and categorize important concepts of a given text into a set of predefined classes. Three main labels are commonly used: persons, locations and organizations [22]. NER techniques can be gathered in two groups: rule-based and machine learning methods. For rule-based methods, the rules are mainly defined manually. They are related to linguistic descriptions, trigger words and lexica of proper names. These rules use patterns and regular expressions in order to locate and classify named entities. Machine learning approaches, on the other hand, aim to extract rules automatically based on learning systems trained on large corpora. Rule-based methods are clearly affected by OCR errors and are not able to deal with the degradation generated by the OCR, whereas, machine learning methods present a sufficient flexibility to be automatically adapted to process noisy texts. More recently, neural networks have been shown to outperform other supervised algorithms for NER. The first deep neural network based learning system has been developed in 2011 [4]. It reached very competitive results for NER in comparison to previous machine learning systems. Therefore, many NER systems using neural networks have been proposed and have shown their abilities to outperform all previous systems [25]. We present in this paper a comparative study of well-performing NER methods. We have chosen, in this work, to use four majors systems available: the well-known NER tool using Conditional Random Fields CoreNLP [8] and three neural network systems BLSTM-CRF [17], BLSTM-CNN [3] and BLSTM-CNN-CRF [20]. The reason being that processing degraded texts using rule based systems require substantial manual efforts to face all typical OCR degradations, unlike machine learning systems which are able to automatically overcome OCR degradations. Furthermore, most rule-based systems are domain-specific or language-dependent and cannot easily be extended to other domains or other languages [9]. Our goal is to evaluate the impact of OCR error on NER accuracy when dealing with noisy text, a task strongly related to document indexing in digital libraries. To the best of our knowledge, no other research work has systematically studied the impact of OCR on named entity recognition over datasets in multiple languages.

[1] Gallica is the digital portal of the National Library of France.

In order to assess our work, we used three publicly available datasets which cover three languages (English, Dutch and Spanish). Given the lack of OCRed annotated data aligned with its ground truth, we have simulated test data by adding typical textual degradation given by an OCR engine. These data have been obtained by automatically adding many levels of degradation in those corpora. More specifically, we spread four types of common OCR degradation in the original clean text. As OCR error depends on the quality and the parameters of the digitization process, we also simulated typical scanning noises at two different levels: rare and reasonably frequent. We finally aligned clean and OCRed data in order to be able to use the same annotation data. Running NER systems through progressively noisy data allows us to draw a graph of NER results relative to OCR error rates. Results over our simulated OCRed resource show a general consistency with a real-life OCRed dataset extracted from Finnish historical newspaper provided by the national library of Finland, which confirms the relevance of our analysis.s

The rest of the paper is organized as follows : Section 2 presents related work studying the impact of OCR. Section 3 consists in an overview of the datasets, followed by outlines of NER results over clean and OCRed texts in Section 4. Section 5 reports our experiments with real data and Section 6 concludes the paper.

2 Related work

Despite decades of research, the output of OCR systems remains imperfect, especially when the original document is old, damaged or poorly digitized. OCR systems lie in the beginning of the digitalization pipeline and OCR errors tend to have a cumulative impact over the subsequent steps. For this reason, researchers have studied the impact of processing text data from noisy sources in order to understand the effects of OCR on text analysis tools.

Much research to process noisy data [32] has stemmed from the field of natural language processing (NLP). Lopresti Daniel [18] for instance considered a text analysis pipeline consisting of sentence boundary detection, followed by tokenization and POS tagging. They reported that among the errors generated by the OCR process, insertion errors were worse than character deletion errors on the sentence boundaries task, while OCR substitution errors were more impactful on POS tagging. The effects of noisy texts have been evaluated also on other NLP tasks such as document summarization [15] and machine translation [36].

Many other works focused on information retrieval from noisy data [5]. Chiron *et al.* [2] proposed a method to estimate the impact of OCR errors on the use of digital libraries. They built an OCR error model using a large corpus of OCRed documents aligned with their corresponding ground truth. Their model allowed the estimation of the risk that a user's query might fail to match with the targeted documents. Taghva *et al.* [33] showed that moderate OCR error rates have not desperate impact on the effectiveness of classical information retrieval measures. Other studies focused on the impact of OCR errors on the

classification of pathology reports for cancer notification [37]. They concluded that OCR errors even with modest rates are not imperceptible for extracting cancer notification items.

For NER, several works have been done to extract NEs from diverse text types such as outputs of Automatic Speech Recognition (ASR) systems [7] informal SMS and noisy social network posts [29]. Palmerand Ostendorf [23] for example described an approach for improving named entity extraction from ASR systems outputs by explicitly modeling errors through the use of confidence scores. In a similar setting, Miller *et al.* [21] have studied the performance of named entity extraction under a variety of spoken and OCRed data. They trained the IdentiFinder system [1] on both clean and noisy input material, performance degraded linearly as a function of word error rates. They concluded that results may lose about 8 points of F-score with only 15% of word error rate. Rodriquez *et al.* [30] reported that manual correction of OCR output have not a very observable improvement on NER results. In [28], Riedl *et al.* presented a complete framework for named entity recognition for both contemporary clean and historical noisy German using transfer learning technique. They achieved state-of-the-art performance for historical datasets with less samples that contains noise. More recently, Hamdi *et al.* [13] and Pontes *et al.* [26] used synthetic OCRed English resources to respectively study the impact of OCR errors on named entity recognition and named entity linking.

In this paper, similarly to [30] and [21], we propose to study the evolution of the performance of named entity recognition systems over noisy OCR data. Unlike them we use more sophisticated NER systems relying on the most recent neural networks models. We also use larger corpora covering four languages, thanks to a technique that allows us to synthesize and test different types and levels of noise. They contain different types of degradation that correspond to the results of long storage and the impact of digitization processes. We defined two levels of degradation for each type in order to obtain a clearer view on OCR errors and their impact on the task of named entity recognition.

3 Dataset overview

To the best of our knowledge, no publicly available corpus has been found with named entity annotations on both clean and noisy texts at the same time. In addition, there are corpora where text produced by an OCR process is aligned with the original text but NEs are not annotated. For this reason, we have taken advantage of three available NER corpora and simulated from them several OCRed versions with variable OCR error rates. We used the public corpora (CoNLL-02 and CoNLL-03) dealing with named entities and covering three languages: English [34], Spanish and Dutch [6]. English data consist of Reuters news stories between August 1996 and August 1997. The Spanish corpus is a collection of news wire articles made available by the Spanish EFE News Agency while the Dutch corpus consists of four editions of the Belgian newspaper "De Morgen". Those datasets are split into three subsets: a training set, a test set and a de-

velopment set. The latter has been built in order to tune parameters of learning methods. All data files contain a single word per line with its associated named entity tag. Table 1 outlines details about each dataset used in this work.

		sentences	words		named entities	
			tokens	terms	tokens	terms
Spanish	Train	8,323	264,715	26,099	32,795	6,821
	Dev	1,915	52,923	9,646	7,567	2,377
	Test	1,517	51,533	9,086	6,178	1,974
English	Train	14,987	204,567	23,624	29,450	6,955
	Dev	3,466	51,578	9,967	7,335	2,735
	Test	3,684	46,666	9,489	7,194	2,384
Dutch	Train	15,806	202,932	27,805	14,555	4,332
	Dev	2,895	37,762	8,151	2,751	1,033
	Test	5,195	68,995	11,803	4,170	1,567

Table 1. CoNLL-02 and CoNLL-03 datasets

The annotation of named entities follows the IOB-scheme (Inside, Outside, Beginning) where every token is labeled as B if the token is the beginning of a named entity, I if it is inside but not the first token within the named entity, or O otherwise [27]. Four classes have been used to label NEs: PER for persons, LOC for locations, ORG for organisations and MISC for other NEs.

From test data, we simulated several OCRed versions. To do so, we first extracted raw texts from test sets and converted them into images. These images have been contaminated by adding typical synthesised noise. We then extracted OCRed data using the Tesseract open source OCR engine v-3.04.01[2] which provides a language package covering many languages among them English, Dutch and Spanish. The subsequent noisy OCRed text and the original one were finally aligned and annotations of the original corpus were projected back on the noisy version. Figure 1 describes the main steps to simulate noisy corpora. We assume that the target text is similar to the indexed text in digital libraries.

In order to contaminate images, we used the DocCreator tool[3] developed by Journet *et al.* [16]. The tool provides many options to add degradation to document images such as blurring, ink degradation and adding phantom characters. In this work, we applied four types of degradation related to storage conditions or poor quality of printing materials that may be present in digital libraries material:

– **character degradation** simulates degradation due to the age of the document or the use of a scanner incorrectly set. It consists in adding small ink spots on characters and can induce the partial obscuration of characters.

[2] `https://github.com/tesseract-ocr/`
[3] `http://doc-creator.labri.fr/`

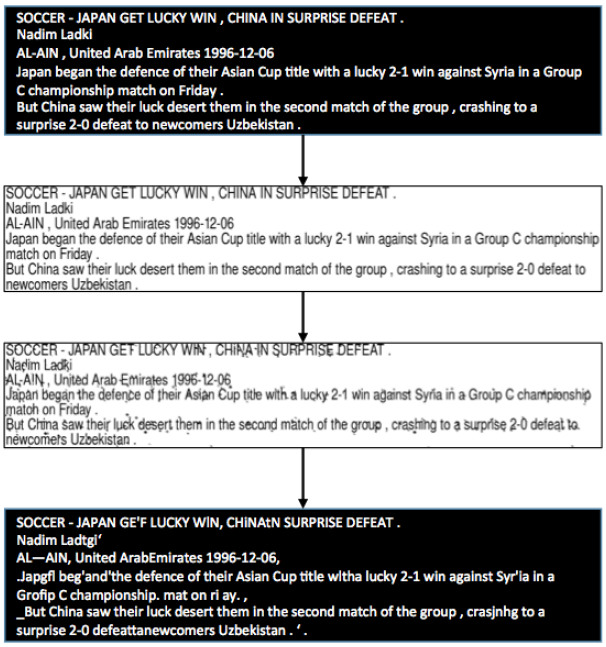

Fig. 1. Simulation of OCRed copora

- **phantom degradation** simulates degradation in worn documents. Following successive uses, some characters can be progressively eroded. The digitization process generates phantom ink around characters.
- **bleed-through** simulates back side ink seeping through the front side of a page. This degradation only appears with double-sided pages.
- **blurring** simulates a blurring effect, as can be encountered during a typical digitization process with focus issue.

For each type of noise, we defined two levels of degradation: LEV-1 where noises are applied rarely and LEV-2 where degradation is reasonably more frequent. These levels allow generating noisy texts with an OCR error rate close to real cases [14]. These degradation levels and types allowed building eight versions for each test corpus. We additionally defined two versions that we call respectively LEV-0 and LEV-MIX. The LEV-0 version is the re-OCRred version of original images with no degradation added while the LEV-MIX version is the result of combining all the LEV-1 degradation types[4]. The LEV-0 degradation aims to evaluate the OCR engine through sharp images whereas the goal of using the LEV-MIX degradation is to be more similar to real-world documents. Degraded documents typically contain several OCR degradations simultaneously.

[4] We have not defined a level combining the LEV-2 degradations because it produces a very poor-quality images and provides unreadable documents.

Following the text extraction by the OCR, the noisy text has been aligned to its original version using the tool RETAS [35]. An example of alignment made between the ground truth and its OCRed version is shown in Figure 2. This alignment reflects the various errors made by the OCR engine. The difference between the two texts is denoted by the presence of the character '@'. Each '@' in the ground truth indicates the insertion of one character by the OCR while '@' in the noisy text indicates that one character has been deleted from the original text.

Fig. 2. Original and noisy texts alignment

In order to evaluate the OCR quality, we used two measures: the Character Error Rate (CER) [14] which corresponds to the proportion of erroneous characters compared to the original text. and the Word Error Rate (WER) [19] which calculates the proportion of erroneous words compared to the total number of words in the original text. A word is considered as erroneous if it contains at least one character error. Table 2 details the OCR error rates at the character and the word levels in the different OCRed version of the three datasets.

As can be seen from Table 2, CER and WER considerably increase when noise is added, comparing to re-OCRed clean text (LEV-0). The table also shows that the noise distributed in the documents is homogeneous. The CER is quite low while the WER is relatively high. Except for Blurring LEV-2 degradation, the CER varies between $\sim$1% and $\sim$7% while the error rate at the word level always exceeds 8%. OCR error rates also show that blurring and character degradation are the most critical noise for digitized documents; they generated the highest error rate both at the character and word levels.

Despite applying the same degradation through all data, OCR is considerably more accurate through Spanish data, CER and WER rates respectively remain below 20% and 30%. On the other hand OCR error rates over English and Dutch data have more variable rates that can reach up to 50%. The bleed-through and phantom characters have a slight impact on the effectiveness of the OCR while ink degradation and blurring lead to the highest OCR error rates. Among these

		English		Dutch		Spanish	
		CER	WER	CER	WER	CER	WER
LEV-0		1.7	8.5	1.6	7.8	0.7	4.8
Bleed-through	LEV-1	1.8	8.5	1.7	8.2	0.8	4.9
	LEV-2	1.8	8.6	1.8	8.9	0.8	5.4
Blurring	LEV-1	6.3	20.0	5.9	22.0	3.0	12.0
	LEV-2	41.3	54.0	27.0	44.7	19.5	29.9
Char deg.	LEV-1	3.6	21.8	4.5	25.1	2.1	14.2
	LEV-2	4.3	23.7	6.4	31.6	2.7	16.3
Phantom deg.	LEV-1	1.7	8.8	1.6	8.0	0.8	5.5
	LEV-2	1.8	10.0	1.7	8.4	0.9	5.9
LEV-MIX		6.9	22.8	5.8	22.2	3.5	11.9

Table 2. Estimation of OCR errors rates

types of degradation, blurring is the most critical degradation that impacted the OCR outputs.

Concerning NEs, knowing their locations in the original text, we aligned them with corresponding words generated by the OCR. We identified then contaminated NEs and those well recognized by the OCR. A total of $3,623$ English named entity tokens have been well recognized by the OCR which represents 63.33%. This rate achieves 72.14% for Spanish and 59.87% for Dutch. All the dataset used in this work are publicly available[5]. We provide for each test corpus (English, Dutch and Spanish) the degraded images and their noisy texts extracted by the OCR as well as the aligned version with clean data at the word and the character levels.

4 Evaluation and results

Neural networks and the related training process require several hyper-parameters such as character embedding dimension, character-based token embedding, LSTM dimension, token embedding dimension, etc. The same parameters for training and testing have been applied on the different dataset: OCRed corpora and clean ones. English embedding has been done using Glove [24] while word2vec [11] was used for Dutch and Spanish word embeddings. Table 3 shows the results of NER on clean datasets. We used traditional metrics ([P]recision, [R]ecall and [F1]-score) to evaluate NER systems.

This first test shows that the results obtained with various methods are globally equivalent for the three languages. We can notice that neural network based approaches give slightly better results than CoreNLP. The same experiments have been run on OCRed dataset. Unsurprisingly, NER accuracy drops proportionally to the rate of OCR errors which is related to the degradation type and level. Table 4 gives the F-score of each NER system on noisy data. Results show

[5] `https://zenodo.org/record/3877554`

		BLSTM-CRF	BLSTM-CNN	BLSTM-CRF-CNN	CoreNLP
English	P	89.54	90.57	91.05	86.35
	R	90.81	90.98	90.75	83.88
	F1	90.17	90.77	90.90	85.10
Dutch	P	79.68	78.61	81.22	74.61
	R	80.96	82.18	79.04	73.28
	F1	80.31	80.36	80.12	73.94
Spanish	P	87.23	87.05	87.54	75.06
	R	83.47	83.21	83.46	76.60
	F1	85.31	85.09	85.45	75.82

Table 3. NER Results on clean data

that compared to clean data, NER results may lose from 3 to 5 points for LEV-0 OCR-ed data. This proves that OCR has a negative impact for the NER task since LEV-0 represents OCR-ed data with no noise added. In other words, even with perfect storage and digitization, NER accuracy may be affected by the OCR quality. For other types of degradation, levels of OCR error rates vary from 8% to 50% at the word level and the NER F-score may drop from 90% to 50% for English. Compared to CoreNLP, deep-learning systems showed a better ability to overcome OCR errors. They achieved satisfactory results when the word error rate was less than 20%.

Results in Table 4 also indicate that the best NER F1-score (in **bold**) can be given by different NER systems according to the type and the level of degradation. For this reason, we calculated the δ measure which gives the minimum decrease rate between the best F1-score given in clean data and the best F1-scores given in noisy data for each type and level of degradation. This measure represents the perfect system that will give the best accuracy for all degradation levels. For the three languages, δ exceeds 40% in noisy data with WER and CER rates reaching more than 0.4 and 0.5 respectively. The Dutch F-score for example decreases under 50% using any one of the four systems through noisy texts extracted from blurred images with an OCR error rate of 44% at the word level.

Figure 3 shows the evolution of the δ measure with respect to degradation. Types of degradation have been sorted according to OCR rates. CER and WER curves are also given for comparison.

5 Experiments on historical dataset

An additional experiment was performed on real life OCRed data, based on Finnish-language historical newspapers from the National Library of Finland (NLF) [31]. The corpus contains around 450K tokens with more than 30K NEs. The NLF corpus distinguishes only two types of NEs: sPER and LOC. The CER and WER rates in the OCRed corpus are respectively 6.96% and 16.67% which is comparable to error rates given by the simulated bleed_LEV-2 degradation in

English	BLSTM-CRF	BLSTM-CNN	BLSTM-CRF-CNN	CoreNLP
Clean	90.17	90.77	**90.90**	85.10
LEV-0	86.77	86.93	**87.45**	79.61
Bleed_LEV-1	85.15	85.08	**86.11**	75.72
Bleed_LEV-2	84.63	**84.72**	83.96	75.27
Blur_LEV-1	**71.03**	70.99	**71.03**	63.39
Blur_LEV-2	59.77	58.98	**60.31**	49.15
DegChar_LEV-1	73.14	**74.22**	74.11	58.12
DegChar_LEV-2	**70.85**	69.43	68.77	55.06
PhantChar_LEV-1	85.59	85.67	**87.01**	74.21
PhantChar_LEV-2	84.58	85.03	**85.20**	73.66
LEV-MIX	**70.87**	70.11	70.82	63.35

Dutch	BLSTM-CRF	BLSTM-CNN	BLSTM-CRF-CNN	CoreNLP
Clean	80.31	**80.63**	80.12	73.94
LEV-0	73.96	73.66	**74.03**	68.36
Bleed_LEV-1	72.10	**73.49**	73.15	66.88
Bleed_LEV-2	72.06	**72.75**	**72.75**	65.45
Blur_LEV-1	63.55	63.56	**63.77**	50.88
Blur_LEV-2	42.78	42.18	**44.56**	30.50
DegChar_LEV-1	57.42	**57.89**	56.33	47.83
DegChar_LEV-2	**51.22**	50.98	50.78	39.16
PhantChar_LEV-1	72.23	**73.66**	73.18	67.12
PhantChar_LEV-2	70.12	**72.99**	72.97	64.15
LEV-MIX	64.33	64.17	**64.88**	53.78

Spanish	BLSTM-CRF	BLSTM-CNN	BLSTM-CRF-CNN	CoreNLP
Clean	85.31	85.09	**85.45**	75.82
LEV-0	85.11	84.25	**85.13**	74.44
Bleed_LEV-1	**84.08**	83.47	84.07	70.15
Bleed_LEV-2	**75.66**	74.99	75.12	68.77
Blur_LEV-1	68.77	66.14	**68.79**	62.41
Blur_LEV-2	60.12	56.73	**61.44**	51.32
DegChar_LEV-1	64.78	63.74	**64.93**	58.33
DegChar_LEV-2	63.01	62.09	**64.12**	52.67
PhantChar_LEV-1	77.12	74.59	**77.21**	68.99
PhantChar_LEV-2	67.77	74.15	**76.76**	67.37
LEV-MIX	72.75	71.17	**73.98**	61.14

Table 4. NER F1-score of noisy data

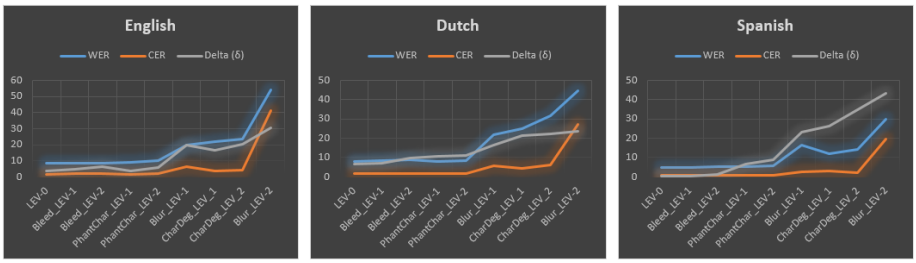

Fig. 3. NER F-score degradation according to OCR error rates

the CoNLL corpora. Ruokolainen *et al.* [31] evaluated the NER annotation of the NLF corpus using CoreNLP. The system respectively yielded overall F1-scores of 71.92% and 78.79% for PER and LOC over OCRed texts which represents a loss of around 9-10 percentage points compared to clean texts. This decrease is mostly equivalent to that obtained on the OCRed synthetic data using CoreNLP (see Table 4). With the same OCR error rates, NER F1-score on the English corpus presents a loss of 9.83% compared to results on clean corpus.

Using BLSTM-CRF, NER F1-score achieves 89.8% and 87.4% on clean and OCRed data respectively which represents a decrease rate of 2.4 percentage points. The corresponding rates in the CoNLL corpora are between 4 and 8 percentage points as shown in Figure 3. Finnish results are slightly better than those obtained with synthetic data using BLSTM-CRF. This is not unexpected since the Finnish training set is larger than the CoNLL datasets. In addition the set of NEs in the NLF corpus is less refined than the set used in the CoNLL corpora. As we showed in Table 4, neural network based systems outperform CoreNLP, we have thus reported the same experiment on the NLF corpus using BLSTM-CRF. Results are shown in Table 5.

		LOC	PER	TOT
clean	P	93.39%	87.43%	90.82%
	R	91.86%	84.68%	88.74%
	F1	92.62%	86.03%	89.77%
OCRed	P	89.68%	83.31%	86.97%
	R	91.06%	83.54%	87.83%
	F1	90.36%	83.42%	87.40%

Table 5. Results on the NLF corpus

Results using the neural-network system are largely outperforming CoreNLP performances. For clean data, we obtained an overall F1-score of 89% (to be compared to 82%). More importantly, for OCRed data, the NER F1-score reaches 90.4% for PER and 83.4% for LOC, resulting in an improvement of around 11

points for both types of NEs. Despite the complexity of the NER task and the occurrence of several types of errors in the documents, the systems achieved interesting results. This proves that they can be used to distinguish named entities in degraded documents. Some word correction strategies, such as auto-encoders, language models, and so on, could be used to decrease the impact of OCR degradation on NER.

6 Conclusion

This paper is the most systematic evaluation of the impact of OCR errors on NER systems over multilingual datasets. We evaluated four machine-learning systems over three available datasets in English, Dutch and Spanish. We re-OCRed these collections and added four types of noises at two different levels in order to simulate various OCR output. All the noisy texts have been aligned with their corresponding ground truth in order to test the NER system through noisy data and to observe the evolution of their accuracy. This new dataset was made publicly available to the community. Such resources, combining OCRed data aligned with their clean version, are very useful for two reasons. First they can be used to train NLP algorithms over collections of documents that have been through an OCR process, as is notably the case of historical documents. Second, they can be used to estimate the impact of OCR over NLP applications and lead to recommendations, for instance on what application can reasonably be run over a document collection given its OCR quality.

We have studied the correlation between OCR error rates and NER accuracy using four effective systems. We showed that NER accuracy drops from 90% to 50% when the word error rate increases from 8% to 50%. These experiments were validated on a real OCR dataset in Finnish, where our systematic study allowed us to outperform the best-known results by $\sim$11 percentage points.

This work showed that specific post OCR correction should be developed in order to improve NER results, and thus improve information access for end users.

Acknowledgements. This work has been supported by the European Union Horizon 2020 research and innovation programme under grant 770299 (NewsEye).

References

1. Bikel, D.M., Schwartz, R., Weischedel, R.M.: An algorithm that learns what's in a name. Machine learning **34**(1-3), 211–231 (1999)
2. Chiron, G., Doucet, A., Coustaty, M., Visani, M., Moreux, J.P.: Impact of ocr errors on the use of digital libraries: towards a better access to information. In: Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries. pp. 249–252. IEEE Press (2017)
3. Chiu, J.P., Nichols, E.: Named entity recognition with bidirectional lstm-cnns. arXiv preprint arXiv:1511.08308 (2015)

4. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of Machine Learning Research **12**(Aug), 2493–2537 (2011)
5. Croft, W., Harding, S., Taghva, K., Borsack, J.: An evaluation of information retrieval accuracy with simulated ocr output. In: Symposium on Document Analysis and Information Retrieval. pp. 115–126 (1994)
6. F, E., Sang, T.K.: Introduction to the conll-2002 shared task: Language-independent named entity recognition. In: Proceedings of CoNLL-2002. pp. 155–158 (2002)
7. Favre, B., Béchet, F., Nocéra, P.: Robust named entity extraction from large spoken archives. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. pp. 491–498. Association for Computational Linguistics (2005)
8. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd annual meeting on association for computational linguistics. pp. 363–370. Association for Computational Linguistics (2005)
9. Gali, K., Surana, H., Vaidya, A., Shishtla, P.M., Sharma, D.M.: Aggregating machine learning and rule based heuristics for named entity recognition. In: Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages (2008)
10. Gefen, A.: Les enjeux épistémologiques des humanités numériques. Socio-La nouvelle revue des sciences sociales pp. 61–74 (2014)
11. Goldberg, Y., Levy, O.: word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722 (2014)
12. Grishman, R., Sundheim, B.: Message understanding conference-6: A brief history. In: COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics. vol. 1 (1996)
13. Hamdi, A., Jean-Caurant, A., Sidere, N., Coustaty, M., Doucet, A.: An analysis of the performance of named entity recognition over ocred documents. In: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL). pp. 333–334. IEEE (2019)
14. Holley, R.: How good can it get? analysing and improving ocr accuracy in large scale historic newspaper digitisation programs. D-Lib Magazine **15**(3/4) (2009)
15. Jing, H., Lopresti, D., Shih, C.: Summarizing noisy documents. In: Proceedings of the Symposium on Document Image Understanding Technology. pp. 111–119 (2003)
16. Journet, N., Visani, M., Mansencal, B., Van-Cuong, K., Billy, A.: Doccreator: A new software for creating synthetic ground-truthed document images. Journal of imaging **3**(4), 62 (2017)
17. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360 (2016)
18. Lopresti, D.: Optical character recognition errors and their effects on natural language processing. International Journal on Document Analysis and Recognition (IJDAR) **12**(3), 141–151 (2009)
19. Lund, W.B., Kennard, D.J., Ringger, E.K.: Combining multiple thresholding binarization values to improve ocr output. In: Document Recognition and Retrieval XX. vol. 8658, p. 86580R. International Society for Optics and Photonics (2013)
20. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional lstm-cnns-crf. arXiv preprint arXiv:1603.01354 (2016)

21. Miller, D., Boisen, S., Schwartz, R., Stone, R., Weischedel, R.: Named entity extraction from noisy input: speech and ocr. In: Proceedings of the sixth conference on Applied natural language processing. pp. 316–324. Association for Computational Linguistics (2000)
22. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes **30**(1), 3–26 (2007)
23. Palmer, D.D., Ostendorf, M.: Improving information extraction by modeling errors in speech recognizer output. In: Proceedings of the first international conference on Human language technology research. pp. 1–5. Association for Computational Linguistics (2001)
24. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
25. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018)
26. Pontes, E.L., Hamdi, A., Sidere, N., Doucet, A.: Impact of ocr quality on named entity linking. In: International Conference on Asian Digital Libraries. pp. 102–115. Springer (2019)
27. Ramshaw, L.A., Marcus, M.P.: Text chunking using transformation-based learning. In: Natural language processing using very large corpora, pp. 157–176. Springer (1999)
28. Riedl, M., Padó, S.: A named entity recognition shootout for German. In: Proceedings of ACL. pp. 120–125. Melbourne, Australia (2018), `http://aclweb.org/anthology/P18-2020.pdf`
29. Ritter, A., Clark, S., Etzioni, O., et al.: Named entity recognition in tweets: an experimental study. In: Proceedings of the conference on empirical methods in natural language processing. pp. 1524–1534. Association for Computational Linguistics (2011)
30. Rodriquez, K.J., Bryant, M., Blanke, T., Luszczynska, M.: Comparison of named entity recognition tools for raw ocr text. In: KONVENS. pp. 410–414 (2012)
31. Ruokolainen, T., Kettunen, K.: À la recherche du nom perdu–searching for named entities with stanford ner in a finnish historical newspaper and journal collection. In: 13th IAPR International Workshop on Document Analysis Systems (2018)
32. van Strien, D., Beelen, K., Ardanuy, M.C., Hosseini, K., McGillivray, B., Colavizza, G.: Assessing the impact of ocr quality on downstream nlp tasks (2020)
33. Taghva, K., Borsack, J., Condit, A.: Effects of ocr errors on ranking and feedback using the vector space model. Inf. Process. Manage. **32**(3), 317–327 (1996)
34. Tjong Kim Sang, E.F., De Meulder, F.: 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. pp. 142–147. Association for Computational Linguistics (2003)
35. Yalniz, I.Z., Manmatha, R.: A fast alignment scheme for automatic ocr evaluation of books. In: Document Analysis and Recognition (ICDAR), 2011 International Conference on. pp. 754–758. IEEE (2011)
36. Yaser, A.O.: Effect of degraded input on statistical machine translation. In: 2005 Symposium on Document Image Understanding Technology. p. 103 (2005)
37. Zuccon, G., Nguyen, A.N., Bergheim, A., Wickman, S., Grayson, N.: The impact of ocr accuracy on automated cancer classification of pathology reports. In: HIC. pp. 250–256 (2012)