[Short paper] When to use OCR post-correction for named entity recognition?

Vinh-Nam Huynh[1], Ahmed Hamdi[2], and Antoine Doucet[2][0000−0001−6160−3356]

[1] University of Science and Technology of Hanoi, ICTLab, Vietnam
namhv.b7114@st.usth.edu.vn
[2] University of La Rochelle, France
{firstname.lastname}@univ-lr.fr

Abstract. In the last decades, a huge number of documents has been digitised, before undergoing optical character recognition (OCR) to extract their textual content. This step is crucial for indexing the documents and to make the resulting collections accessible. However, the fact that documents are indexed through their OCRed content is posing a number of problems, due to the varying performance of OCR methods over time. Indeed, OCR quality has a considerable impact on the indexing and therefore the accessibility of digital documents. Named entities are among the most adequate information to index documents, in particular in the case of digital libraries, for which log analysis studies have shown that around 80% of user queries include a named entity. Taking full advantage of the computational power of modern natural language processing (NLP) systems, named entity recognition (NER) can be operated over enormous OCR corpora efficiently. Despite progress in OCR, resulting text files still have misrecognised words (or noise for short) which are harming NER performance. In this paper, to handle this challenge, we apply a spelling correction method to noisy versions of a corpus with variable OCR error rates in order to quantitatively estimate the contribution of post-OCR correction to NER. Our main finding is that we can indeed consistently improve the performance of NER when the OCR quality is reasonable (error rates respectively between 2% and 10% for characters (CER) and between 10% and 25% for words (WER)). The noise correction algorithm we propose is both language-independent and with low complexity.

Keywords: named entity recognition, optical character recognition, character degradation, spelling correction

1 Introduction

Large quantities of valuable documents have been scanned as images for digital archives. In order to extract text information from those images, OCR techniques are widely used. The OCR process usually begins with loading text images as input and improving the input quality, a step that may involve multiple techniques, such as deskewing, noise removal, etc. In the next steps, OCR systems binarize images and detect text zones. Then, the core part of OCR systems will take place by mapping each character image to the most proper character code. Finally, the OCR system generates a text file corresponding to the input image. However, due to storage conditions or poor quality of printing materials, the image quality may be low, in which case the OCR may generate very noisy texts, strongly diverging from the original text, known as the Ground Truth (GT). Often, these noisy texts are nonetheless readable by humans in digital libraries, which lessens the motivation to re-digitize and/or re-OCR them, which is a costly process. The key problem is that this noisy text is used for building the indexes used for instance by search engines. This implies that a keyword query will return documents containing the adequate keyword only if it was properly recognized by the OCR system. Many relevant documents may thus be missed, in proportions that are very hard to quantify.

A study has shown that named entities are the first point of entry for users in a search system [4]. For instance, on the Gallica digital library[3], 80% of user queries contain at least one named entity [1]. For this reason, named entities can be given a higher semantic value than other words to index digitised documents. In order to improve the quality of user searches in a system, it is thus necessary to ensure the quality of these particular terms. In the presence of OCR errors, NER systems are not able to override the degradation caused by the OCR in the extracted text. For this reason, post-OCR task should be helpful in order to improve the effectiveness of NER systems over noisy textual data.

This work extends a previous work studying the performance of an effective neural network-based NER system over several noisy versions of a NER corpus with variable rates of OCR errors [5]. We aim to use a post-OCR correction to this variety of OCRed texts in order to quantitatively estimate its contribution on NER performance. The underlying idea of this work is to evaluate the impact of post-OCR correction on the performance of NER over noisy text, a task strongly related to information access in digital libraries.

The remainder of this paper is organized as follows: Section 2 surveys related works on misspellings, OCR errors and post-OCR approaches. Then, we introduce the dataset in Section 3. In Section 4, we analyze OCR errors and give many useful statistics, before summarizing our major findings in Section 5.

2 Related Work

Many studies focused on the impact of OCR errors on NLP [8] and Information Retrieval (IR) [19]. Miller *et al.* [12], for instance studied the performance of named entity extraction under a variety of spoken and OCRed data. They showed that over noisy texts, NER F-score may lose about 8 points with a word error rate of only 15%. Recently, Hamdi et al. [5] simulated many noisy versions of NER resources with different types of noise in order to study the correlation between OCR error rates and NER accuracy. In a similar setting, [17] studied

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

the impact of OCR errors on different NLP tasks. They concluded that NER is less affected by OCR errors than sentence segmentation or dependency parsing.

A few amount of works studied the contribution of post-OCR correction on NLP and IR tasks. Magdy and Darwish [11], for instance, examined the effect of OCR error correction on document retrieval. On named entity recognition, Rodriquez *et al.* [16], reported that manual correction of OCR output have not a very observable improvement on NER results.

Our work is similar to Rodriquez *et al.* [16], we study the impact of post-OCR correction on NER performance. However, unlike them, we automatically rectify erroneous tokens over a variety of noisy texts using a low-complexity algorithm. We perform NER using three accurate neural network NER systems.

3 Dataset overview

The dataset used in this work is the English corpus given by the Conference on Natural Language Learning in 2003 (CoNLL-2003) [18]. The dataset defines four classes of named entities: **Persons** (PER) including individuals and groups. **Locations** (LOC) includes countries, regions, addresses as well as states and provinces. **Organisations** (ORG) concerns commercial, educational, government as well as medical-science, religious, sports. **Miscellaneous** (MISC) annotates all other named entities such as nationalities and events. The dataset defines more then 40,000 named entities.

As we mentioned in the introduction, this work extends a previous study on the impact of OCR errors on named entity recognition [5]. Authors simulated several OCRed versions[4] of the test data adapted to this real-life problem. This simulation of document degradation is required because while there exist datasets with OCRed text and corrected text, as well as text with NER markup, there are no datasets contain both, and even less so with different levels and types of OCR noise. First of all, raw texts in the test set have been extracted and then converted into images. With the help of the DocCreator tool [6], common OCR degradation have been added to these images by putting noise texture to their backgrounds. Degradation include bleeding effect, blurring, character degradation, and phantom character. For each type of noise, two levels of degradation were applied: level 1 corresponds to noises that are sparsely applied on the original document and level 2 corresponds to noises that appear more often. Thus, level 1 of each degradation means the simulated text contains less noise than level 2. These degradations define typical OCR noises when storing documents in digital libraries or using a document scanner [3]. The open source OCR engine Tesseract v-3.04.01 has been used to extract noisy texts from the degraded images. In order to quantify OCR error rates in the obtained versions of the test set, two common metrics have been used: the character error rate (CER) and the word error rate (WER) which correspond respectively to the rate of erroneous output characters (resp. words) out of the total number of

[4] https://zenodo.org/record/3877554

characters (resp. words) in the corpus [9]. In the end, many OCRed versions of the test set are obtained with a CER and WER rates respectively varying from 1% to 20% and from 8% to 50%.

4 Named Entity Recognition on noisy texts

For NER, we utilized the DeLFT[5] (Deep Learning Framework for Text) framework. This library re-implements standard state-of-the-art deep learning architectures relevant to named entity recognition. Among the existing architectures, we chose to use BLSTM ones due to their ability to overcome some of the OCR errors [15]. We built three models based on BLSTM-CRF [7], BLSTM-CNN [2] and BLSTM-CNN-CRF [10]. We also used the Stanford Global Vectors (GloVe) as our word embedding in order to represent document vocabulary and word features. GloVe is an unsupervised learning algorithm that produces a word vector space based on global word co-occurrence statistics [14].

Results show comparable NER performances of the three systems. However, they are harmfully impacted by the OCR quality especially when the OCR error rates are relatively high. Figure 1 shows the correlation between the NER performances and the character error rate. We show also the evolution of the word error rate (dotted line). Regardless of the system used NER results may fall by about 30 percentage points due to OCR noise when the OCR error rates are respectively 20% and 50% at the character and the word levels. Unsurprisingly, the higher the OCR error rates, the greater the degradation of NER F1-score. For all systems, the NER F1-score achieves less than 80% when the CER reaches around 3% and the WER is about 20%.

As proof, noisy texts contain many out-of-vocabulary words, which NER models cannot identify as named entities. Following our analysis of the output predictions, we made several observations:

1. Contaminated named entities were well recognised by NER system in both clean and noisy versions: for instance, the named entity *Mittermayer*, which corresponds to a person name is correctly associated to *Minermayer*. However it is well recognized and labeled by the NER system.
2. Contaminated named entities were detected and well classified in the clean text version, but their alternative in noisy version were wrongly recognized by the NER system: the location *Japan* for example is associated to *Japgfl* which is not recognized by any NER system.
3. Named Entities that were not corrupted after the OCR process still failed to be recognized by NER systems, because of noisy context surrounding them.

[5] https://github.com/kermitt2/delft

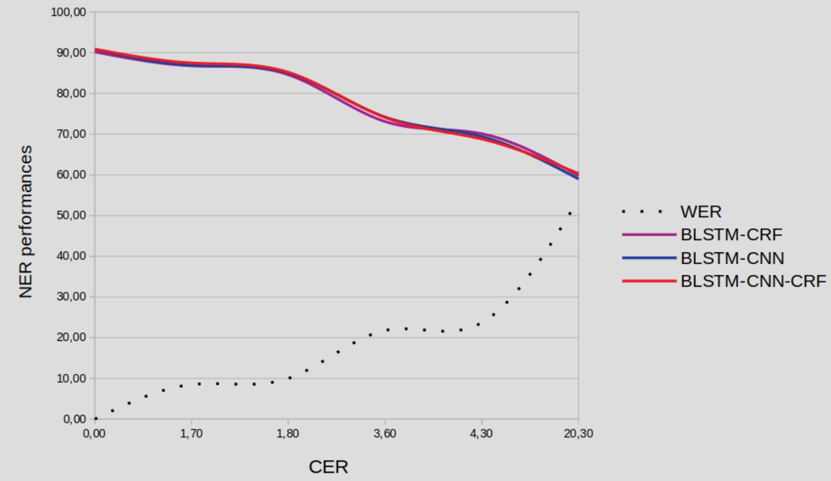
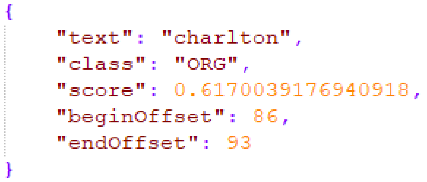


Fig. 1. NER F1-score degradation according character error rates



```
{
    "text": "charlton",
    "class": "ORG",
    "score": 0.6170039176940918,
    "beginOffset": 86,
    "endOffset": 93
}
```

Fig. 2. Correctly-OCRed named entity were wrongly classified by NER system due to noisy paragraph surround

The example in Figure 2 shows that even if the named entity is not contaminated by the OCR system, it can be impacted by noisy surrounding words and therefore associated to a wrong class.

- GT: ”Prime Minister Dick Spring who said the honour had been made in recognition of **Charlton** ’s achievements as the national soccer manager.”
- OCR: ”Prime Minister Dick Spring who said the homur had been made in recognimm on **Charlton** s ac ievements as the national soccer manager.”

Since most of these noisy words had the same problem, the existing NER systems wrongly classified or did not recognize them. Hence, noisy texts dramatically reduced NER performance.

In order to solve this problem, we proposed to use an edit distance based algorithm as a spell-checker including a dictionary in order to carefully examine each word and compare it with every dictionary entry.

5 Experiments

To handle this problem, we pre-processed noisy texts before parsing them using NER models with an efficient and low-complexity text correction method named SymSpell[6]. In our work, the algorithm is based on Levenshtein distance which calculates the minimum steps (insertion, deletion or substitution) to transform a string into another string.

5.1 Noisy text correction

SymSpell consists of two steps: pre-calculation and searching. At first, SymSpell generates all possible terms within the pre-set edit distance by deletion only. In this work, the max edit distance is set to 2. According to a recent study, OCR post-processing approaches are recommended to focus on correcting erroneous words with edit distances 1 and 2 [13]. For example, with (italy, 2) meaning the word "italy" and a max edit distance of 2, we have:

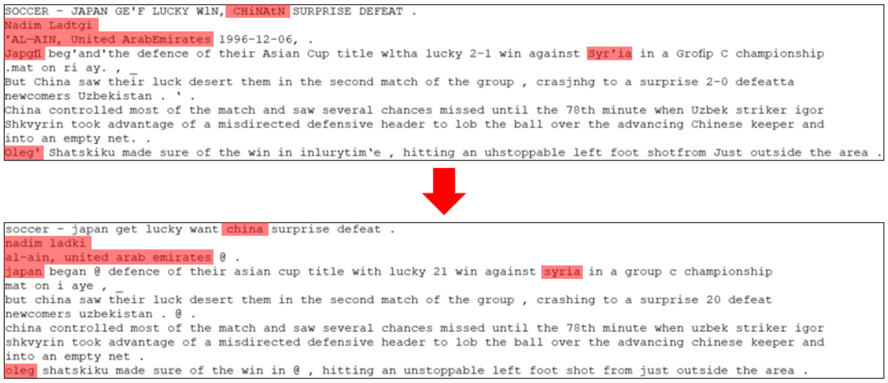
- delete (italy, 0) == italy
- delete (italy, 1) == ital or itay or taly or ... (for a total of $\binom{5}{1}$ possible strings)
- delete (italy, 2) == ita or itl or aly or ... (for a total of $\binom{5}{2}$ possible strings)
- Many different entries may share the same result string: delete (italy, 1) == delete (vital, 1) == ital

Second, upon receiving input, SymSpell starts to erase each single character within an edit distance from that term. By doing so, both imprecise dictionary generated strings and imprecise input-generated string might match and meet in the middle. Thus, SymSpell will choose possible candidates and give suggestions to correct the misspelled input. Setting an edit distance threshold allows SymSpell to remove many irrelevant candidates. SymSpell automatically chooses the one with the highest frequency when it encounters multiple candidates that satisfy the max edit distance threshold.

5.2 Results and discussion

Figure 3 presents the output of SymSpell text pre-processing. Well corrected named entities are colored in red.

[6] https://github.com/mammothb/symspellpy



```
SOCCER - JAPAN GE'F LUCKY WlN, CHiNAtN SURPRISE DEFEAT .
Nadim Ladtgi
'AL-AIN, United ArabEmirates 1996-12-06, .
Japgfl beg'and'the defence of their Asian Cup title wltha lucky 2-1 win against Syr'ia in a Grofip C championship
.mat on ri ay. , _
But China saw their luck desert them in the second match of the group , crasjnhg to a surprise 2-0 defeatta
newcomers Uzbekistan . ' .
China controlled most of the match and saw several chances missed until the 78th minute when Uzbek striker igor
Shkvyrin took advantage of a misdirected defensive header to lob the ball over the advancing Chinese keeper and
into an empty net. .
Oleg' Shatskiku made sure of the win in inlurytim'e , hitting an uhstoppable left foot shotfrom Just outside the area .
```

```
soccer - japan get lucky want china surprise defeat .
nadim ladki
al-ain, united arab emirates @ .
japan began @ defence of their asian cup title with lucky 21 win against syria in a group c championship
mat on i aye , _
but china saw their luck desert them in the second match of the group , crashing to a surprise 20 defeat
newcomers uzbekistan . @ .
china controlled most of the match and saw several chances missed until the 78th minute when uzbek striker igor
shkvyrin took advantage of a misdirected defensive header to lob the ball over the advancing chinese keeper and
into an empty net .
oleg shatskiku made sure of the win in @ , hitting an unstoppable left foot shot from just outside the area .
```

Fig. 3. SymSpell correction output

In order to evaluate the contribution of SymSpell on NER results, we calculated NER F1-scores of different models before and after the post-OCR correction. Table 1 compares the F1-scores given on the original noisy data and the ones that applied the SymSpell method.

OCR error rate		BidLSTM-CRF		BidLSTM-CNN		BidLSTM-CRF-CNN	
CER	WER	Original	SymSpell	Original	SymSpell	Original	SymSpell
1.7	8.5	**86.8**	79.8	**86.9**	78.9	**87.6**	80.0
1.7	8.8	**85.6**	79.6	**85.7**	78.9	**87.0**	80.0
1.8	8.0	**84.6**	79.4	**85.0**	78.8	**85.2**	79.8
1.8	8.5	**85.2**	79.7	**85.0**	78.9	**86.1**	80.0
1.8	8.6	**84.6**	79.8	**84.7**	78.8	**84.0**	80.0
3.6	20.0	73.1	**78.6**	74.2	**77.6**	74.1	**78.2**
4.3	21.8	70.9	**78.7**	69.4	**77.6**	68.8	**78.7**
6.3	23.7	71.0	**78.0**	71.0	**77.6**	71.0	**77.2**
20.3	54.0	59.8	**68.4**	59.0	**68.3**	60.3	**68.0**

Table 1. F1-score comparison between original noise and SymSpell correction

Table 1 shows that the BLSTM-CNN-CRF model globally outperforms the two other NER models in both OCRed and post-OCR corrected texts. The post-OCR correction improves NER results when the OCR error rates are relatively high. Very satisfactory results (up to 77%) are reached when the word error rate is less than 25%. However, post-OCR correction may also degrade NER F1-scores especially when the OCR error rate is very low (less than 2% at the character level and less than 10% at the word level). The SymSpell method did not take care of surrounding context and relied only on pure edit-distance, then chose the most suitable word by frequency index, the algorithm sometimes changed original words into ones that were not related to the context (e.g., substituted ”Al-ain” - a location NE - with the word ”Again”). This mechanism would reduce the performance of existing NER system mentioned above.

In order to stress the impact of the OCR noise and the contribution of the post-OCR correction on NER F1-scores, we calculated two δ measures:

- δ_{noisy} which gives the decrease rate between the F1-score given in clean data and the F1-score given in noisy data using BLSTM-CNN-CRF.
- $\delta_{symSpell}$ which indicates the decrease rate between the F1-score given in clean data and the F1-score given in post-OCR corrected data using BLSTM-CNN-CRF.

Figure 4 shows the evolution of the δ measures according to the character error rate. The WER curve (dotted) is also given to ease comparison.

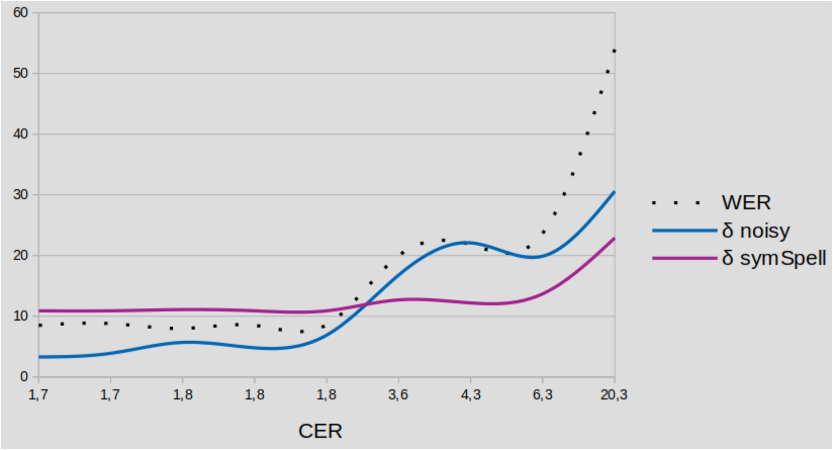


Fig. 4. F1-score decrease on noisy and corrected text

The curves show that NER F1-scores are considerably impacted by OCR errors when the OCR error rate is relatively high (more than 20% at the character level and more than 50% at the word level). The decrease rate δ_{noisy} jumps from 20% to 30%. The post-OCR correction allowed us to overcome this issue. $\delta_{symSpell}$ is almost constant $\sim$ 10% which means that SymSpell allowed us to overcome OCR issues and propose a NER F1-score exceeding 80%. However, Figure 4 also showed that for low error rates (less than 2% and 10% at the character level and the word level respectively), a post-OCR correction is not a suitable solution and it is better to simply run NER systems on the original noisy text, as if they contained no noise.

6 Conclusion

The main aim of this research was to propose methods that help increase the performance of NER over noisy text data, by applying post-OCR correction. The result has shown that the SymSpell correction algorithm (with max edit distance set to 2) can consistently increase NER results over noisy texts when the CER and the WER respectively exceed 2% and 10%, while standard techniques are otherwise preferable. In future work, we plan to further study this phenomenon, using different max edit distances and exploiting other post-OCR correction techniques.

Acknowledgment

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

References

1. 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)
2. Chiu, J.P., Nichols, E.: Named entity recognition with bidirectional lstm-cnns. arXiv preprint arXiv:1511.08308 (2015)
3. Farahmand, A., Sarrafzadeh, H., Shanbehzadeh, J.: Document image noises and removal methods (2013)
4. Gefen, A.: Les enjeux épistémologiques des humanités numériques. Socio-La nouvelle revue des sciences sociales (4), 61–74 (2014)
5. 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)
6. 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)
7. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360 (2016)
8. 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)
9. 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)
10. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional lstm-cnns-crf. arXiv preprint arXiv:1603.01354 (2016)
11. Magdy, W., Darwish, K.: Effect of ocr error correction on arabic retrieval. Information Retrieval **11**(5), 405–425 (2008)

12. 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)
13. Nguyen, T.T.H., Jatowt, A., Coustaty, M., Nguyen, N.V., Doucet, A.: Deep statistical analysis of ocr errors for effective post-ocr processing. In: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL). pp. 29–38. IEEE (2019)
14. 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)
15. 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
16. 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)
17. van Strien, D., Beelen, K., Ardanuy, M., Hosseini, K., McGillivray, B., Colavizza, G.: Assessing the Impact of OCR Quality on Downstream NLP Tasks:. In: Proceedings of the 12th International Conference on Agents and Artificial Intelligence. pp. 484–496. SCITEPRESS - Science and Technology Publications, Valletta, Malta (2020). https://doi.org/10.5220/0009169004840496, http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0009169004840496
18. 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)
19. 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)