An Analysis of the Performance of Named Entity Recognition over OCRed Documents

Ahmed Hamdi, Axel Jean-Caurant, Nicolas Sidere, Mickaël Coustaty, Antoine Doucet
L3i Laboratory, University of La Rochelle
La Rochelle, France
{firstname.surname}@univ-lr.fr

ABSTRACT

The use of digital libraries requires an easy accessibility to documents which is strongly impacted by the quality of document indexing. Named entities are among the most important information to index digital documents. According to a recent study, 80% of the top 500 queries sent to a digital library portal contained at least one named entity [2]. However most digitized documents are indexed through their OCRed version which includes numerous errors that may hinder the access to them.

Named Entity Recognition (NER) is the task that aims to locate important names in a given text and to categorize them into a set of predefined classes (person, location, organization). This paper aims to estimate the performance of NER systems through OCRed data. It exhaustively discusses NER errors arising from OCR errors; we studied the correlation between NER accuracy and OCR error rates and estimated the cost of character insertion, deletion and substitution in named entities. Results show that even if the OCR engine does contaminate named entities with errors, NER systems can overcome this issue and correctly recognize some of them.

KEYWORDS

Indexing, OCR, Named Entity, extraction, digital libraries

1 INTRODUCTION

In digital libraries, large quantities of printed documents are scanned and archived as images. Text extraction using Optical Character Recognition (OCR) is therefore necessary for indexing documents which is an essential feature for the accessibility to these documents. Unfortunately, the quality of OCR output is not always good and still sometimes far from the ground truth. Comparing to the costly efforts that can be spent on fixing OCR errors, it is considered that the quality of OCR outputs is sufficient for humans to read and explore documents. However, several research works suggest that the effectiveness of systems processing OCR output might be considerably harmed by OCR errors [9].

A study has shown that named entities are the first point of entry for users in a search system [5]. In order to improve the quality of user searches in digital libraries, it is thus necessary to ensure the quality of these particular terms. NER approaches have emerged in the 1990's [6], and the early systems relied on Rule-based approaches. Rules used in those systems are defined by humans and based on dictionaries, trigger words and linguistic descriptors. Such techniques require a lot of time and effort to be extracted and handled. Thus they cannot be easily updated to new types of texts or entities. To overcome this problem, the efforts on NER are largely dominated by Machine Learning techniques including sequential tagging methods such as Hidden Markov Models [1], and Conditional Random Fields (CRFs) [4]. Since 2011, deep learning methods have emerged and topped NER evaluations [3].

In the presence of OCR error, rule-based methods are unable to overcome the degradation generated by an imperfect OCR process. On the other hand, machine learning methods present a sufficient flexibility to be adapted to processing noisy text. For this reason, we used in this work the neural network system based LSTM-CRF [8] to recognize named entities both on clean and noisy texts. It generates the most probable sequence of predicted labels from surrounding words. Long Short-Term Memory (LSTM) networks compute a representation of the context of each word. A CRF layer allows generating the most probable sequence of predicted labels from surrounding words.

The underlying idea is to evaluate the impact of OCR errors on NER accuracy and to analyze NER errors arising from OCR errors when dealing with noisy texts, which is strongly related with the indexing process of documents in digital libraries.

2 METHODOLOGY AND RESULTS

Since there exists no OCRed annotated NER data aligned with its ground truth, we took advantage of an existing NER corpus of clean text [10] and generated OCR noise. The corpus consists of Reuters news stories between August 1996 and August 1997. The training set contains more than $264k$ tokens containing $32k$ named entities whereas the test set has about $51k$ words and $6,178$ named entities. From the test set, we simulated noisy text data by adding typical textual degradation caused by an OCR engine. More specifically, we extracted raw text from the test set and converted it into images. These images have been contaminated used the DocCreator tool[1] developed by Journet et al.[7]. The tool allowed to add four common types of OCR degradation related to storage conditions or poor quality of printing materials that may be present on digital libraries: **Character degradation** adds small ink spots on characters in order to simulate degradation due to the age of the document or the use of an incorrectly set scanner. **Phantom degradation** simulates eroded characters that can appear in worn documents following successive uses. **Bleed-through** is typical of double-sided pages. It simulates the ink from the back side that seeps through the front side. Lastly, DocCreator provides an option to simulate a **blurring effect**. We further extract noisy text from OCRed data using the Tesseract open source OCR engine v-3.04.01. Original and noisy texts are finally aligned and gold annotations are projected back to the noisy versions. We assume that the target text is similar to the indexed text in digital libraries.

[1] http://doc-creator.labri.fr/

This process allows to build four OCRed versions from the original dataset. We additionally define a fifth version (called **LEV-0**) that corresponds to re-OCRred version of original images with no degradation added. It aims to evaluate the OCR engine through sharp images. The same parameters for training and testing have been applied on both the OCRed and clean versions of the datasets.

Table 1 shows NER results on clean and OCRed texts, with respect to OCR error rates: additionally to Character Error Rate (CER) and Word Error Rate (WER), we calculated the proportion of erroneous named entities extracted by the OCR system (ENER).

	OCR			NER		
	CER	WER	ENER	Pre	Rec	F1-score
Clean	−−	−−	−−	89.4	90.8	90.1
LEV-0	1.7	8.5	6.9	83.7	90.7	86.8
Bleed	1.8	8.6	7.1	84.0	84.1	84.1
PhantChar	1.7	8.8	7.8	75.8	78.6	77.1
Blurring	6.3	20.0	21.5	66.9	69.5	68.8
CharDeg	3.6	21.8	23.4	64.5	64.8	64.7

Table 1: NER performance over noisy data, for undegraded OCR (LEV-0), bleed-through (Bleed), phantom degradation (PhantChar), Blurring effect and character degradation (CharDeg)

For instance, we can see in Table 1 that for LEV-0, the precision of NER results is lowered by 5.7 points with CER of only 1.7%. This proves 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 20% at the word level and NER results through these data can drop from 90% to 60%.

Experiments showed that even if named entities are wrongly extracted by the OCR, NER systems can overcome OCR errors and recognize part of the NEs correctly especially when error rates are not too high (cf. Figure 1).

GT:	Pierre	Van	Hoydonk	scored	for	Glasgow	Rangers
gold:	PER	PER	PER	O	O	LOC	LOC
OCR Lev_0:	Dierre	Van	Hoyoonk	scored	for	Glasgow	Ramgers
GT:	**PER**	**PER**	O	O	O	**ORG**	**ORG**
OCR Blur:	Pierre	|n	Hoyoom	scores	on	Glasgow	Rimfen
GT:	**PER**	O	O	O	O	*LOC*	O

Figure 1: Named Entity Recognition through OCRed data

Following the application of Blur degradation for instance, 21.5% of named entities are incorrectly extracted by the OCR; among them 8.1% are properly recognized by the NER system. However, the latter did not succeed to overcome OCR errors on around 81% of contaminated named entities especially when their Levenshtein distance from the ground truth exceeds 1. This figure is equal 37.9% when the distance is 1 and 68.9% when the distance is 2 while it rises to 100% with longer distances. The named entity "*Madrid*" for instance, is recognized as "*Madnid*" and "*Nudaid*" by the OCR after applying respectively the LEV-0 and the Blur degradations. Although "*Madnid*" is correctly extracted and labeled, "*Nudaid*" is not considered as named entity. The error analysis also shows that 77% of the named entities that are wrongly recognized contain at least one character insertion, that 73% of them contain at least one substitution and that 68% of them contain at least one deletion.

3 CONCLUSION

In this paper we simulated many OCR outputs by adding different noises to the CONLL-03 NER corpus. The experiments simulate a common use case of digital libraries, in which documents are indexed through an OCRed version. We evaluated the evolution of NER accuracy depending on the level of noise present in the text. We concluded that NER accuracy drops from 90% to 60% when the WER and CER rates in the OCR outputs increase from 1% to 7% and from 8% to 20% respectively. From these rates, we assume that current state-of-the-art NER algorithms can only be relied upon when the OCR quality is sufficiently good. Given the large share of search queries containing NEs that are submitted to digital libraries, we believe these results are very important to the community.

In future works, we plan to add weights to OCR outputs at the character and the word levels. Using knowledge on the share of erroneous characters and words, we could assign probabilities to predicted named entities and improve the precision of NER systems. Another important point that we aim to study is NE-focused OCR post-correction. The important focus on NEs of digital library users implies that post-OCR solutions strictly focused on NEs would have a high impact on information access.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Daniel M Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1998. Nymble: a high-performance learning name-finder. *arXiv preprint cmp-lg/9803003* (1998).

[2] Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, Muriel Visani, and Jean-Philippe Moreux. 2017. 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*. IEEE Press, 249–252.

[3] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12, Aug (2011), 2493–2537.

[4] Michele Filannino, Gavin Brown, and Goran Nenadic. 2013. ManTIME: Temporal expression identification and normalization in the TempEval-3 challenge. *arXiv preprint arXiv:1304.7942* (2013).

[5] Alexandre Gefen. 2014. Les enjeux épistémologiques des humanités numériques. *Socio-La nouvelle revue des sciences sociales* 4 (2014), 61–74.

[6] Ralph Grishman and Beth Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, Vol. 1.

[7] Nicholas Journet, Muriel Visani, Boris Mansencal, Kieu Van-Cuong, and Antoine Billy. 2017. DocCreator: A New Software for Creating Synthetic Ground-Truthed Document Images. *Journal of imaging* 3, 4 (2017), 62.

[8] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360* (2016).

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

[10] Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*. Association for Computational Linguistics, 142–147.