Impact of OCR errors on the use of digital libraries Towards a be er access to information

Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, Muriel Visani, Jean-Philippe Moreux

Impact of OCR errors on the use of digital libraries

Towards a better access to information

Guillaume Chiron
National Library of France
Quai François Mauriac
Paris, France 75706
guillaume.chiron@bnf.fr

Antoine Doucet
L3i Lab, University of la Rochelle
Avenue Michel Crépeau
La Rochelle, France 17042 Cedex 1
antoine.doucet@univ-lr.fr

Mickaël Coustaty
L3i Lab, University of la Rochelle
Avenue Michel Crépeau
La Rochelle, France 17042 Cedex 1
mickael.coustaty@univ-lr.fr

Muriel Visani
L3i Lab, University of la Rochelle
Avenue Michel Crépeau
La Rochelle, France 17042 Cedex 1
muriel.visani@univ-lr.fr

Jean-Philippe Moreux
National Library of France
Quai François Mauriac
Paris, France 75706
jean-philippe.moreux@bnf.fr

ABSTRACT

Digital collections are increasingly used for a variety of purposes. In Europe only, we can conservatively estimate that tens of thousands of users consult digital libraries daily. The usages are often motivated by qualitative and quantitative research. However, caution must be advised as most digitized documents are indexed through their OCRed version, which is far from perfect, especially for ancient documents. In this paper, we aim to estimate the impact of OCR errors on the use of a major online platform: The Gallica digital library from the National Library of France. It accounts for more than 100M OCRed documents and receives 80M search queries every year. In this context, we introduce two main contributions. First, an original corpus of OCRed documents composed of 12M characters along with the corresponding gold standard is presented and provided, with an equal share of English- and French-written documents. Next, statistics on OCR errors have been computed thanks to a novel alignment method introduced in this paper. Making use of all the user queries submitted to the Gallica portal over 4 months, we take advantage of our error model to propose an indicator for predicting the relative risk that queried terms mismatch targeted resources due to OCR errors, underlining the critical extent to which OCR quality impacts on digital library access.

KEYWORDS

Digital libraries, OCR errors, indexation bias, search logs

ACM Reference format:
Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, Muriel Visani, and Jean-Philippe Moreux. 2017. Impact of OCR errors on the use of digital libraries. In *Proceedings of ACM/IEEE-CS Joint Conference on Digital Libraries, Toronto, Ontario, Canada, June 2017 (JCDL'17),* 4 pages.
DOI: 00.000/000_0

1 INTRODUCTION

The growing use of digital libraries along with the upcoming qualitative and quantitative new usages bring more than ever the indexation and the retrieval processes in the spotlight. The accuracy of Optical Character Recognition (OCR) technologies considerably impacts the way digital documents are indexed, consulted and exploited [16]. Offering relevance and exhaustiveness to users' needs thus remains a major issue [7]. During the last decades, OCR engines have been constantly improved and are nowadays able to produce exploitable results on mainstream documents. But in practice, digital libraries contain many transcriptions with quality below expectations, especially when it comes to ancient documents that often include challenging layouts and various levels of conservation [2]. Let us mention this emblematic use case of Gallica[1]: the word "budget" is often transcribed as "gadget" in 19^{th} century press documents, long before that word even existed, which is clearly problematic. These kinds of errors, given the scale at which digital libraries are queried, are propagated in the document processing pipeline and lead to vast consequences. Knowing the magnitude of the phenomenon and understanding its nature is essential to undertake an appropriate solution to solve the problem. Through this study, we offer raw data as well as related statistics that should help to apprehend the extent of issues related to the indexation bias induced by OCR errors. To this end, the present paper provides two contributions:

- an original 12M characters (2.1 M tokens) corpus of OCRed documents along with the corresponding GS (Gold Standard[2]), with an equal share of English- and French-written documents. Statistics on OCR errors have been computed thanks to a novel alignment method which is also detailed;
- an indicator for predicting the relative risk that queried terms mismatch targeted resources due to OCR errors based on user queries submitted to Gallica over 4 months.

1.1 Related work

Datasets of documents including their OCRed version and their corresponding gold standard have been made available through

[1] Digital library of the National Library of France: http://gallica.bnf.fr
[2] Trustworthy corpora commonly used for developing machine learning algorithms.

978-1-5386-3861-3/17/$31.00 ©2017 IEEE

various projects (*e.g.* IMPACT[3], Europeana Newspapers[4], ISRI[5]). The accuracy of OCR engines has been evaluated either on the basis of such public datasets (not aligned, often not generic and not standardized), on private datasets built on measure (*e.g.* 2M labeled tokens [1]) or on artificially generated ground truths [6] which intrinsically suffers from bias. Although the OverProof project [5] proposes monolingual datasets made of a few million symbols, to our knowledge, no GS collection has ever become an evaluation standard. The dataset we are proposing will be the largest aligned dataset (gold standard and original OCRed data) to be made public which mixes multilingual and multiple sources of various difficulties.

In the domain of OCR quality assessment, numerous studies have tackled the problem of the prior prediction [19] or posterior estimation [15] of the results of an OCR engine at the document, word or character level. This can be used to improve the accuracy of automatic processes and to further provide worthwhile feedback. Although a few studies have highlighted risks inherent to OCR errors on quantitative analysis [16], no estimation of query mismatch in terms of retrieval relative to a given corpora have been done. These specific difficulties of document access in digital libraries have been addressed in two ways, through content- and query-based approaches.

Content-based approaches tackle the problem at the source by improving the content of the digital materials themselves thanks to denoising methods (commonly relying on post-OCR correction). Many ideas have been explored in that direction, such as: modeling the OCR output as a noisy channel to recover the original version [10, 13]; combining outputs of multiple OCR systems [9, 17]; using Google's online suggestions as correction candidates [3]; active learning of human post-editing [1].

Query-based approaches try to handle the problem afterward by making the retrieval process more flexible. Given a query, fuzzy retrieval methods such as query expansion propose to generate extra queries that are more likely to match the intended materials. Different generation rules have been explored in the literature: manually identifying phonetic similarities and common typing errors [4]; pairs of correct and incorrect words (e.g. OCR aligned with the gold standard) [8]; models of bigram words and measures (edit distances) [12].

Both types of approaches cited above are learning-based and require as many statistics on OCR errors as possible, so as to be modeled and trained. In order to allow the research community to dig deeper into the analysis of how OCR errors impact the use of DL, we present our dataset and detail the proposed OCR/GS alignment method (Section 2). New statistics on OCR errors are then presented and combined to Gallica query logs for an in-context investigation on the impact of OCR errors for DLs (Section 3).

2 PROPOSED DATASET

The proposed dataset has been built within the AméliOCR project[6] for research on OCR post-correction. The dataset is made publicly available in the context of the ICDAR2017 Competition on Post-OCR Text Correction.

2.1 Document collection

The collection accounts for 12M OCRed characters along with the corresponding GS, with an equal share of English- and French-written documents (see Table 1). The documents come from different digital collections available, among others, at the National Library of France (BnF) and the British Library (BL). The corresponding GS comes both from BnF's internal projects and external initiatives such as Gutenberg, Europeana Newspapers, IMPACT and Wikisource.

Degraded documents sometimes result in highly noisy OCR output and thus cannot reasonably be fully aligned with their GS. The unaligned sequences have not been included in the presented statistics (e.g. number of characters, error rate). Error rates vary according to the nature and the state of degradation of the documents. Historical newspapers for example, due to their complex layout and their original fonts have been reported to be especially challenging for OCR engines with up to 10% of wrongly detected characters on some documents.

Lang	Source	Type	Dates	E.R.	Char.
Eng.	BL Euro NP	serials	1744 - 1894	4%	1.8 M
	BL Monog	monog.	1858 - 1891	1%	1.2 M
	GT BnF Eng	monog.	1802 - 1911	2%	3.0 M
Fr.	Europeana NP	serials	1814 - 1944	4%	1.0 M
	IMPACT	monog.	1821 - 1864	1%	0.4 M
	GT BnF Fr	mixed	1686 - 1943	1%	2.0 M
	Digit. BnF	mixed	1654 - 2000	3%	0.2 M
	News other	serials	1897 - 1934	4%	0.6 M
	Monog other	monog.	1689 - 1883	3%	1.8 M
				Total:	12 M

Table 1: Sources, quantities and average Error Rates (E.R.) involved in both English and French parts of the dataset.

Our dataset comes from nine different sources, includes documents in two languages, and ranges over four centuries. These features give a wide overview of documents and OCR errors that can be found in a digital library. A comparison of the frequency indexes extracted from our corpus with those of well-known large corpora (i.e. Google 1-gram, Wikipedia, mixed corpus of literature monographs) revealed same orders of magnitude concerning the words distribution, which comforts the idea that our dataset is representative. It is further the only collection for which a GS of 12 million characters with such heterogeneity is provided.

2.2 Gold standard alignment (GSA)

The computation of statistics on OCR errors requires an alignment phase between the OCR output and the GS at character-level. We processed documents of various formats (*e.g.* books, compilations, newspapers), some of them individually containing up to 500k characters. Traditional alignment tools (e.g, [14], or open-source extensions[7]) are not able to deal with such huge corpora. Better optimized approaches (*e.g.* [18]) have recently been proposed and usually rely on LCS (Longest Common Subsequences) as anchors

[3]IMPACT, European Commission's 7th Framework Program, grant agreement 215064
[4]EU Competitiveness and Innovation Framework Programme grant ENP 297380
[5]Data to evaluate OCR accuracy: http://code.google.com/p/isri-ocr-evaluation-tools
[6]Led by the National Library of France (Department of preservation and conservation) and the L3i laboratory (Univ. of La Rochelle, France)
[7]https://github.com/kba/awesome-ocr

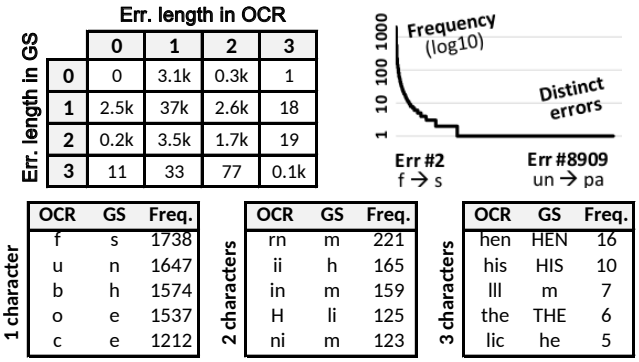
for subdividing long sequences into subsequences until these subsequences can reasonably be processed with classical sequence aligners [11].

However, such strict anchoring mechanisms cause problems for highly noisy OCRed texts because it cannot identify enough similar subsequences to perform a recursive alignment. Thus, for the need of our dataset which contains a large number of highly degraded documents, we propose a similar alignment approach, but instead of relying on strict LCS, our method relies on fuzzy LCS. In short, it works by maximizing the similarity of character distribution among a pair of sliding windows (one is attached to the OCR part, the other to the GS part). The size of the sliding windows is initialized to be as large as possible, yet in coherence with the limits imposed by the downstream sequence aligner, and is recursively down-sized while the matching criteria have not been reached for the given pair of subsequences. Additionally, for optimization purpose, we assume that in most cases the reading order is respected and thus both sliding windows (on the OCR and on the GS) follow a linear progression along the document. When this assumption fails (often the case for documents with a complex layout), a search mechanism, optimized by a dichotomous approach, takes over and re-synchronizes the sliding windows on a more probable location. Finally, this proposed alignment approach offers a trade-off between the efficiency of the strict LCS search process and tolerance to noise.

2.3 Details on OCR errors

Based on the GSA, we estimate alphanumerical OCR errors amount to 52k, with a similar number of affected tokens which corresponds to 2.5% of the 2.1 M tokens found in the corpus. Estimating the number of non-alphanumerical errors is quite complex because of different encodings, inconsistent hyphenations, varying punctuation and spaces. Based on our corpus, Figure 1 shows statistics based on error length between 0 and 3 characters (few errors exceed that length). Moreover, the graphic at the top of the figure shows how OCR errors are distributed, along with a few examples of the most common errors involving 1, 2 and 3 characters in the tables below. A more in-depth study of erroneous tokens led us to the following estimates:

- 15% of the wrongly OCRed terms are Named Entities;
- About 50% of OCR errors are made on terms which do not belong to classical dictionaries (e.g, the OpenOffice dictionary).



Err. length in GS \ Err. length in OCR	0	1	2	3
0	0	3.1k	0.3k	1
1	2.5k	37k	2.6k	18
2	0.2k	3.5k	1.7k	19
3	11	33	77	0.1k

1 character

OCR	GS	Freq.
f	s	1738
u	n	1647
b	h	1574
o	e	1537
c	e	1212

2 characters

OCR	GS	Freq.
rn	m	221
ii	h	165
in	m	159
H	li	125
ni	m	123

3 characters

OCR	GS	Freq.
hen	HEN	16
his	HIS	10
lll	m	7
the	THE	6
lic	he	5

Figure 1: Distribution and frequencies of the 52k OCR errors of our dataset along with sample details.

3 IMPACT ON SEARCH RESULTS FROM THE GALLICA PLATFORM

We noted an important correlation between terms frequencies in Gallica and those of end-users searches. To that end, we collected search logs from the Gallica platform over a period of 4 months (December 2015 to March 2016). It is worth noting that since most users of the Gallica platform are French speakers, the query logs naturally contain mostly French terms. A set of 28M user queries was collected, set aside from another 49M queries which were considered as having been performed by robots[8]. Highly redundant queries (*i.e.* up to 10k occurrences) have also been filtered out under the assumption that they were submitted by a single user (perhaps an undetected robot), or perhaps stemming from a hyperlink (a query directly available on the Internet). The queries often contain multiple terms (38% are composed of only one word, 23% are composed of two words, 14% of three, and 25% of four words or more). The search log of our study is finally composed of 26k real query terms along with their frequencies.

3.1 Named entities highly exposed

It is important to notice that a large number of queries involve named entities in at least one of their terms (80%, or 400 out of the first 500). But when observing individual frequencies of searched terms in a global manner, we observe that only 30% of the terms are named entities. They mostly correspond to places (e.g. France-68k, Paris-43k, Maroc-33k) and proper names (e.g. Louis-37k, Charles-25k, Eugène-23k). Even if well-known named entities are included in common dictionaries, and can thus reasonably be corrected, the problem remains for the many terms appearing in the long tail. Two thirds of the searched terms have a frequency below 200, which is a much lower order of magnitude than the common terms cited above. They correspond to terms that do not belong to common dictionaries (e.g. “Bodenehr”). Therefore, it becomes interesting to study the proportion of searched terms that do not belong to dictionaries, and to estimate the frequency of OCR errors on those terms. Highly exposed terms that do not belong to common dictionaries shall be especially difficult to correct using post-correction approaches (and to retrieve using a standard retrieval model).

3.2 Impact on common search words

The whole 26k searched terms match 2.5% of the GS (about 300k tokens). Let us focus on terms that were commonly searched (at least 35 times) and for which the system did not provide relevant results. The cross analysis between the OCR errors observed in our corpus and the search logs highlights that among the 300k tokens matched, 8k are affected by OCR errors. That is, 2k common search terms are causing mismatches, meaning that 7% of the queried terms potentially miss documents due to OCR errors.

3.3 Low represented queries in corpora

Built-in dictionaries are often used to automatically enhance search results. It is difficult to find an optimal threshold of term frequency to establish those dictionaries, however infrequent terms of the

[8]Robots were filtered using a common policy on the user-agent, excluding terms such as ”bot”, ”spider” or ”slurp”.

corpora generally tend to not be included. Such infrequent terms are therefore more exposed to OCR errors than the frequent ones.

An important part of highly searched terms (from the query logs) corresponds to terms uncommon in the dataset, such as "brauwer", "guerche" or "djidjelli". Those kind of examples are among the 100 most searched terms while they are only found in a few thousand documents in Gallica, and thus require particular attention.

3.4 Query exposure model

A query having terms lowly represented in the corpora does not necessary mean a high failure exposure. It is important to take the difficulty of the letter patterns into account, with regard to the OCR engines. By taking advantage of our OCR error model (c.f. Figure 1) as well as frequency distribution information, we propose an indicator (1) for estimating the exposure of a term to any search bias. In the following, we assume that low-represented terms have low chances to be corrected by existing methods – due to their absence in common dictionaries – and thus are likely to expose the user to missing targeted resources:

$$risk_t = \frac{1}{log(r_t) * len(t)^2} \sum_{p=1}^{nP} log(1 + f_p) \quad (1)$$

where $risk_t$ is an relative indicator estimating the risk for a term t to fail at recovering intended resources, r_t is the frequency of the term t in the corpora, nP is the number of OCR error patterns identified in the term, $len(t)$ is the length of the term, and f_p the frequency of the error pattern given the model.

3.5 Experiments on real queries

Figure 2 shows the correlation between the proposed risk estimator (equation 1) and the real error ratio (relatively to each term) for the 2k most affected terms searched in Gallica. For example, the terms "quelques" and "toujours" are wrongly OCRed respectively 0.6% and 0.8% of the time. As they are frequent terms and also do not include frequent erroneous letter patterns, their risk exposure is relatively small ($risk_t < 0.05$). However, words like "connell" or "fellowes" are wrongly OCRed respectively 31% and 66% of the time. As they are infrequent terms and contain frequent erroneous patterns, their risk exposure is rather high ($risk_t > 0.15$).

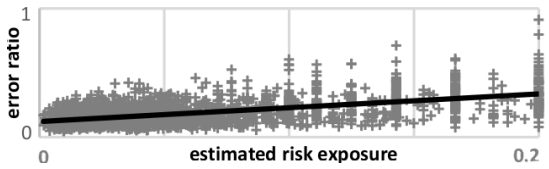


Figure 2: Error ratio of the 2k most affected terms in our corpus, compared to our error exposure indicator.

4 CONCLUSION

In this article we have presented an original corpus of OCRed documents that accounts for 12M characters from 9 sources written in 2 languages. All the original texts have been aligned at the character level with their corresponding GS, using a custom alignment approach based on fuzzy LCS (Section 2.2). This alignment allowed us to build an OCR error model (Figure 1) to support further analysis.

We have studied the correlation between search logs – relying on 4 months of user queries performed on Gallica – and OCR errors found in our dataset. This led us to the definition of an estimator of the risk that a user query fails to recover intended resources (equation 1). Although the results presented in Figure 2 are preliminary, we have put forward the possibility to approximate the likelihood of a given query to result in poor retrieval results, due to OCR errors. We have also shown that a significant amount of user queries are affected by wrongly OCRed terms that do not belong to usual dictionaries. This underlines the potential of post-OCR correction methods that do not strictly rely on lexicons.

In future works, based on our combined dataset (OCRed documents and search logs), we would like to perform a quantitative and qualitative study on the two common families of approaches – OCR post-correction and query expansion – relying on our OCR error model. We also wish that the distribution of our unique dataset will foster further research in the DL research community.

This study is part of the AmélioCR project supported by the BnF's 8^{th} quadrennial research plan (2016-2019).

REFERENCES

[1] Ahmad Abdulkader and Mathew R Casey. 2009. Low cost correction of OCR errors using learning in a multi-engine environment. In *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*. IEEE, 576–580.
[2] Beatrice Alex, Claire Grover, Ewan Klein, and Richard Tobin. 2012. Digitised historical text: Does it have to be mediOCRe?.. In *KONVENS*. 401–409.
[3] Youssef Bassil and Mohammad Alwani. 2012. OCR Post-Processing Error Correction Algorithm Using Google's Online Spelling Suggestion. *Journal of Emerging Trends in Comp. and Info. Sciences* 3 (2012).
[4] Wolfram M Esser. 2004. Fault-tolerant fulltext information retrieval in digital multilingual encyclopedias with weighted pattern morphing. In *European Conference on Information Retrieval*. Springer, 338–352.
[5] John Evershed and Kent Fitch. 2014. Correcting noisy OCR: Context beats confusion. In *Proceedings of the 1st International Conference on Digital Access to Textual Cultural Heritage*. ACM, 45–51.
[6] Shaolei Feng and R Manmatha. 2006. A hierarchical, HMM-based automatic evaluation of OCR accuracy for a digital library of books. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 109–118.
[7] Kripabandhu Ghosh, Anirban Chakraborty, and al. 2016. Improving Information Retrieval Performance on OCRed Text in the Absence of Clean Text Ground Truth. *Information Processing & Management* 52, 5 (2016), 873–884.
[8] David Hawking, Paul Thistlewaite, and Peter Bailey. 1997. ANU/ACSys TREC-5 Experiments. *Recall* 35, 34 (1997), 36.
[9] Shmuel T Klein, M Ben-Nissan, and M Kopel. 2002. A voting system for automatic OCR correction. (2002).
[10] Okan Kolak and Philip Resnik. 2002. OCR error correction using a noisy channel model. In *Proceedings of the second international conference on Human Language Technology Research*. Morgan Kaufmann Publishers Inc., 257–262.
[11] Heng Li and Nils Homer. 2010. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in bioinformatics* 11, 5 (2010), 473–483.
[12] Kwong Bor Ng, David Loewenstern, Chumki Basu, Haym Hirsh, and Paul B Kantor. 1996. Data Fusion of Machine-Learning Methods for the TREC5 Routing Task (and other work).. In *TREC*.
[13] Martin Reynaert. 2014. TICCLops: Text-Induced Corpus Clean-up as online processing system.. In *COLING (Demos)*. 52–56.
[14] Stephen V Rice and Thomas A Nartker. 1996. The ISRI analytic tools for OCR evaluation. *UNLV/Information Science Research Institute, TR-96-02* (1996).
[15] Ahmed Ben Salah, Jean philippe Moreux, Nicolas Ragot, and Thierry Paquet. 2015. OCR performance prediction using cross-OCR alignment. In *Document Analysis and Recognition, 2015 13th International Conference on*. IEEE, 556–560.
[16] Myriam C Traub and al. 2015. Impact analysis of OCR quality on research tasks in digital archives. In *International Conference on Theory and Practice of Digital Libraries*. Springer, 252–263.
[17] Martin Volk, Lenz Furrer, and Rico Sennrich. 2011. Strategies for reducing and correcting OCR errors. In *Lang. Technology for Cultural Heritage*. Springer, 3.
[18] Ismet Zeki Yalniz and Raghavan Manmatha. 2011. A fast alignment scheme for automatic ocr evaluation of books. In *2011 International Conference on Document Analysis and Recognition*. IEEE, 754–758.
[19] Peng Ye and David Doermann. 2012. Learning features for predicting OCR accuracy. In *Pattern Recognition, 21st International Conf. on*. IEEE, 3204–3207.