Assessing the Impact of OCR Errors in Information Retrieval

Guilherme Torresan Bazzo, Gustavo Acauan Lorentz, Danny Suarez Vargas, and Viviane P. Moreira(✉)

Institute of Informatics, UFRGS, Porto Alegre, Brazil
{guilherme.bazzo,gustavo.lorentz,dsvargas,viviane}@inf.ufrgs.br

Abstract. A significant amount of the textual content available on the Web is stored in PDF files. These files are typically converted into plain text before they can be processed by information retrieval or text mining systems. Automatic conversion typically introduces various errors, especially if OCR is needed. In this empirical study, we simulate OCR errors and investigate the impact that misspelled words have on retrieval accuracy. In order to quantify such impact, errors were systematically inserted at varying rates in an initially clean IR collection. Our results showed that significant impacts are noticed starting at a 5% error rate. Furthermore, stemming has proven to make systems more robust to errors.

Keywords: OCR · Retrieval effectiveness · Noisy text

1 Introduction

Estimates say that most information useful for organizations is represented in an unstructured format, predominantly as free text [6]. A significant portion of this useful information is stored in PDF files – research articles, books, company reports, and presentations are typically disseminated in PDF format. PDF documents need to be converted into plain text before being processed by an Information Retrieval (IR) or a text mining system. These files can either be *digitally created* or created from *scanned* documents. While the former are generated from an original electronic version of a document (*i.e.,* contain the text characters), the later contain images of the original document and need to go through Optical Character Recognition (OCR) so that their contents can be extracted. Despite being addressed by researchers for decades, OCR is still imperfect. As a result, the extracted text contains errors that typically involve character exchanges. Although digitally created PDFs are cleaner, they are not problem-free since, for example, hyphenated terms (due to separation into syllables) may be identified as two tokens and indexed incorrectly.

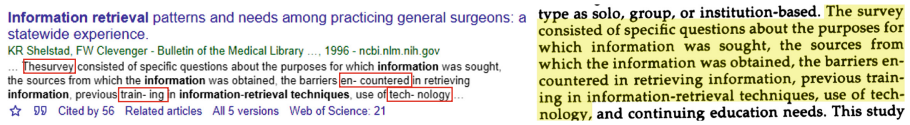
Extraction errors can have a negative impact on the quality of IR systems and are found even in mainstream search engines. Figure 1 presents an excerpt of the result page generated by Google Scholar for the query "information retrieval techniques". In the small snippet from a matching document, we can see four

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 102–109, 2020.
https://doi.org/10.1007/978-3-030-45442-5_13

errors – three terms were erroneously segmented into two tokens, and two terms were concatenated into one token. The effect is that a query with the correct spelling for *e.g.,* "the barriers encountered in retrieving information" would be unable to retrieve that document. Approaches for treating misspelled queries cannot solve this problem as the issue is in the document, not in the query. Furthermore, there are important differences between the types of errors made by humans while typing and those made by OCR systems [9].

The fact that this is still an open issue is evidenced by two recent competitions organized in the scope of the International Conference on Document Analysis and Recognition (ICDAR) [2,12]. The best performing approaches employ state-of-the-art methods such as character-based Neural Machine Translation and recurrent networks (bidirectional LSTMs) taking BERT models as input. The best results for the error detection task were below 0.7 in terms of F1 in several languages [12], showing that there is still a lot of room for improvement.

Our goal is to revisit the problem of retrieving OCR-ed text and quantify the impact that these errors have on the accuracy of IR systems. Ideally, to quantify the impact, one needs an IR test collection with source PDF files, their extracted and corrected versions, a set of queries, and their corresponding relevance judgments. Unfortunately, to the best of our knowledge, such a collection does not exist and creating one would demand significant effort. In line with previous works on this topic [3,7], our approach was to systematically insert errors in a standard IR collection containing plain text documents, queries, and relevance judgments. Different error rates were tested so that we could gauge their effects. In order to simulate real errors, we collected, assessed, and manually corrected a sample of OCR-ed PDF documents. Statistics drawn from this sample were used to guide the error insertion approach. Our experiments were performed in an IR collection containing documents in Portuguese – a language that makes use of diacritics (*e.g.,* à, á, ã, â, é, í, ç *etc.*). These characters are typically among the ones with more extraction errors. The results showed that error rates starting at 5% can cause a significant impact in many system configurations and that stemming makes systems more robust to coping with errors.



(a) Excerpt from GoogleScholar (b) Source PDF file

Fig. 1. Example of extraction errors identified in a mainstream search engine.

2 Related Work

Existing work on dealing with OCR-ed texts spans over a long period and focused on approaches for detecting and fixing errors [4,5,9,10]. Specifically on the topic

of improving the retrieval of OCR text, Beitzel *et al.* [1] surveyed a number of solutions – most of which date to the late 1990s. TREC ran a confusion track to assess retrieval effectiveness on degraded collections. Their modified test collections had 5 and 20% character error rates. Five teams took part in the challenge. The organizers reported that counter-intuitive results had been found and that "there is still a great deal to be understood about the interaction of the diverse approaches" [7]. Croft *et al.* [3] share some similarities with our work. However, rather than injecting errors into a clean text collection, the authors opted to randomly select words to be discarded from the document and, as a consequence, they were not indexed. The limitation of such approach is that it does not account for issues with wrong segmentation (adding or suppressing the space character) or cases in which the error modifies the word into another valid word. The main finding was that performance degradation was more critical for very short documents. In a detailed investigation, Taghva *et al.* [13] observed that while the results seem to have insignificant degradation on average, individual queries can be greatly impacted. Furthermore, they report an impressive increase in the number of index terms in the presence of errors and that relevance feedback methods are unable to overcome OCR errors.

This paper differs from existing works in a number of aspects. The configurations we assess include the use of stemming, more recent ranking algorithms, and more levels of degradation. Finally, we experiment with a different test collection in a language that has not been extensively used for IR.

3 Simulating Errors

The methodology we propose to insert errors is shown in Fig. 2. Our goal is to replicate, as much as possible, the pattern of problems that actually happen in PDF conversions to plain text from both digitally created and scanned documents. In order to achieve that, one needs a sample of aligned pairs of extracted and expected contents (shown as input in Fig. 2). The expected contents need to be manually produced by correcting the extracted text. This is a laborious and time-consuming task. By comparing these <extracted, expected> pairs at character level, we generate a *character exchange list*.

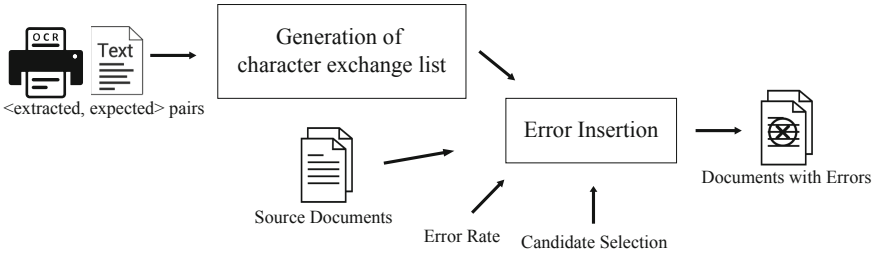


Fig. 2. Approach for error injection in documents.

To align the <extracted, expected> pairs, we used the Needleman-Wunsch [8] algorithm. This algorithm generates the best (global) alignment of two sequences, with the addition of gaps to account for mismatching characters. We found exchanges of one-to-one (*e.g.,* "inserted" → "insorted"), one-to-two (*e.g.,* "document" → "docurnent"), or two-to-one (*e.g.,* "light" → "hght") characters. The frequencies of the exchanges were computed and stored in the character exchange list. Then, they are used to bias the error insertion algorithm towards the most frequent exchanges.

By analyzing the pattern of errors found, we came up with a categorization of the types of issues. (*i*) *Exchange of characters.* This is the most common error found (90% of all errors) and it is caused by the low quality of the documents we are processing. Every exchange in our exchange list has assigned to itself the frequency of its appearance, which we use, in conjunction with the tournament selection, to elect one error to a given term. (*ii*) *Separated terms.* This error corresponds to 5% of the cases and it happens when a space character is erroneously inserted in the middle of a term. (*iii*) *Joined terms.* This error, which has a frequency of 4.9%, happens when the space between terms is omitted, resulting in the unexpected concatenation of terms. (*iv*) *Erroneous symbol.* This issue accounts for 0.1% of all errors, usually represents dirt or a printing error at the scanned document.

Issues (*i*) to (*iii*) can potentially affect recall as relevant documents containing terms with these problems will not be retrieved by the query. Issue (*iv*) can also lower precision since the fragment of a term can match a query for which the document is not relevant (*e.g.,* if the term "encounter" found in a document d is fragmented into the tokens "en" and "counter", then d can erroneously match a query with the term "counter").

Two alternatives for the selection of candidate terms were employed. In the first, any term from any document could be selected. In the second, a more targeted selection was made in which candidate terms were taken only from judged documents (*i.e.,* the documents in the qrels file).

Using the desired error rate, we iterate through every candidate term in the documents. The term is chosen to be modified with a probability equivalent to the given error rate. If the term is selected, then the choice of error is made taking the observed frequency. This was achieved by drawing a random float between 0 and 1 and matching it against the corresponding error frequency. The selection of which exchange to apply was made using tournament selection in ten rounds according to the frequency of the exchange.

4 Experimental Evaluation

This Section describes the experimental evaluation of the error insertion method to assess the impact of OCR errors in IR systems. The resources, tools, and configurations used in our experiments were as follows.

Data. To generate the character exchange list, we took a sample of 900 PDF documents containing abstracts from research articles published at the website

of a Brazilian Oil Company[1]. The extracted text was manually checked and the extraction errors were fixed to create the list of <extracted, expected> pairs. The IR collection used was Folha de São Paulo, a Brazilian Newspaper. It has 103K documents, 100 queries, and it has been used in important evaluation campaigns such as CLEF [11].

Table 1. Number of index terms (in thousands) and the proportional increase in comparison to the baseline.

Setting	Baseline	1%	5%	10%	25%	50%
ALL-NS	273	355 (30%)	523 (91%)	659 (141%)	937 (243%)	1,243 (355%)
ALL_ST	203	253 (24%)	352 (73%)	434 (113%)	605 (197%)	801 (293%)
JD_NS	273	342 (25%)	473 (73%)	574 (110%)	770 (182%)	983 (260%)
JD_ST	203	245 (20%)	324 (59%)	386 (90%)	514 (153%)	660 (224%)

Tools. The OCR software used was Abbyy Finereader 14[2]. The choice was made after its good result compared to a number of other alternatives including Tesseract, a9t9, Omnipage, and Wondershare. The IR System was Apache Solr[3].

Experimental Procedure. In our experimental procedure, we varied the following parameters. The *Ranking function*, taking three possibilities: Cosine (COS) using TF-IDF weighting, BM25, and Divergence from Randomness (DFR). The *use of stemming*: applying a light stemmer (ST) and no stemming (NS). The *error rates* were 1%, 5%, 10%, 25%, and 50%. Baseline runs using the original documents were also created. The *candidate terms* for error insertion were either any term from any document (ALL) or any term from the judged documents (JD). These variations amounted to a total of 72 experimental runs, which were evaluated using standard IR metrics. Statistical significance was measured using T-tests. Queries were made by simply taking the title field from the topics. The goal was to simulate real queries that are typically short.

Table 1 shows the number of index terms for the combination of error rates, use of stemming, and candidate terms. As expected, the number of index entries grows remarkably with the error rates, reaching more than a four-fold increase for unstemmed runs with a 50% error rate.

The results for all experimental runs are in Table 2. The runs in which the mean average precision (MAP) decrease was found to be statistically significant (in relation to the baseline) at a 99% confidence interval are in a darker shade and the ones with a 95% significance are in a lighter shade.

The best ranking function in terms of absolute MAP values was DFR, followed by BM25. However, there were no differences on their robustness in

[1] http://publicacoes.petrobras.com.br/.
[2] https://www.abbyy.com/.
[3] https://lucene.apache.org/solr/.

Table 2. MAP Results for all configurations. The numbers in brackets indicate the proportional change.

Setting	Baseline	1%	5%	10%	25%	50%
ALL-BM25-NS	0.251	0.249 (−0.8%)	0.243 (−3.2%)	0.232 (−7.7%)	0.211 (−16.2%)	0.156 (−38.1%)
ALL-BM25-ST	0.294	0.291 (−0.8%)	0.288 (−1.9%)	0.281 (−4.4%)	0.263 (−10.4%)	0.223 (−23.9%)
ALL-COS-NS	0.242	0.243 (0.2%)	0.239 (−1.3%)	0.218 (−9.8%)	0.202 (−16.5%)	0.166 (−31.6%)
ALL-COS-ST	0.275	0.273 (−0.9%)	0.266 (−3.6%)	0.256 (−7.1%)	0.251 (−8.8%)	0.223 (−19.0%)
ALL-DFR-NS	0.263	0.262 (−0.2%)	0.258 (−1.8%)	0.240 (−8.8%)	0.222 (−15.6%)	0.180 (−31.5%)
ALL-DFR-ST	0.307	0.304 (−1.1%)	0.306 (−0.4%)	0.295 (−4.2%)	0.276 (−10.1%)	0.250 (−18.7%)
JD-BM25-NS	0.251	0.248 (−1.4%)	0.243 (−3.4%)	0.232 (−7.6%)	0.226 (−10.2%)	0.169 (−32.9%)
JD-BM25-ST	0.294	0.290 (−1.2%)	0.284 (−3.4%)	0.280 (−4.6%)	0.261 (−11.2%)	0.229 (−21.9%)
JD-COS-NS	0.242	0.240 (−0.6%)	0.237 (−2.0%)	0.218 (−10.0%)	0.200 (−17.3%)	0.152 (−37.1%)
JD-COS-ST	0.275	0.277 (0.7%)	0.271 (−1.7%)	0.263 (−4.5%)	0.247 (−10.4%)	0.209 (−24.2%)
JD-DFR-NS	0.263	0.263 (0.0%)	0.252 (−4.2%)	0.239 (−9.2%)	0.221 (−16.2%)	0.163 (−38.2%)
JD-DFR-ST	0.307	0.306 (−0.4%)	0.300 (−2.3%)	0.294 (−4.3%)	0.281 (−8.6%)	0.236 (−23.2%)

the presence of OCR errors as their pattern of MAP decrease was the same. Strangely, in two runs in which the cosine was used, the insertion of errors at a 1% rate improved the performance (ALL-COS-NS and JD-COS-ST). This can be explained by the fact that errors are inserted both relevant and non-relevant documents. In these cases, the errors were introduced in non-relevant documents which made relevant documents be ranked higher.

The use of stemming consistently improved the results – *i.e.,* all runs in which stemming was used had higher scores than their unstemmed counterparts. Stemming has made the runs more robust to the OCR errors. This can be seen comparing the loss in MAP of the runs with and without stemming. Nearly all runs in which stemming was used had smaller losses than their counterparts. Furthermore, the aid of stemming is more noticeable in the runs with higher error rates. The benefit of stemming can be explained by the fact that the OCR error can be in the suffix that is removed. Looking at the correlation between the number of index terms (Table 2) and MAP, we find a strong negative correlation of 0.86. When the correlation is measured for stemmed and unstemmed runs separately, the negative correlations are 0.81 and 0.90, respectively. This gives further support to the benefits of stemming.

Looking at our sample of aligned extracted and expected texts (assembled from real documents) we observed an error rate of 1.5%. Considering this rate and the results in Table 2, one can conclude that the errors do not have a severe impact on IR as significant impacts are observed starting at 5%. At a 10% rate, all runs are significantly affected. Recall that this small error rate was found using the software which provided the best results on relatively recent documents. Some studies that provide statistics of the proportion of errors found in OCR-ed documents report finding error rates of around 20% in historical documents [4,5,14]. At that error rate, the degradation is considered statistically significant.

Comparing the two choices of candidate terms for error insertion we find close scores. This means that the error injection targeting the judged documents did not have an influence on the results.

5 Conclusion

Despite having been investigated for decades, the issues associated with retrieving noisy text still remain unsolved in many IR systems. In this paper, we revisit this topic by assessing the impact that different error rates have on retrieval performance. We tested different setups, including ranking algorithms and the use of stemming. Our findings suggest that statistically significant degradation starts at a word error rate of 5% and that stemming is able to make systems more resilient to these errors. As future work, it would be useful to assess which type of error identified in (Sect. 3) has the greatest impact in retrieval quality.

Acknowledgments. This work was partially supported by Petrobras, CNPq/Brazil, and by CAPES Finance Code 001.

References

1. Beitzel, S.M., Jensen, E.C., Grossman, D.A.: A survey of retrieval strategies for OCR text collections. In: Proceedings of the Symposium on Document Image Understanding Technologies (2003)
2. Chiron, G., Doucet, A., Coustaty, M., Moreux, J.: ICDAR 2017 competition on post-OCR text correction. In: International Conference on Document Analysis and Recognition (ICDAR), vol. 01, pp. 1423–1428 (2017)
3. Croft, W.B., Harding, S., Taghva, K., Borsack, J.: An evaluation of information retrieval accuracy with simulated OCR output. In: Symposium of Document Analysis and Information Retrieval (1994)
4. Droettboom, M.: Correcting broken characters in the recognition of historical printed documents. In: Proceedings 2003 Joint Conference on Digital Libraries, pp. 364–366, May 2003
5. Evershed, J., Fitch, K.: Correcting noisy OCR: context beats confusion. In: Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage (DATeCH 2014), pp. 45–51 (2014)
6. Grimes, S.: Unstructured data and the 80 percent rule, p. 10. Carabridge Bridgepoints (2008)
7. Kantor, P.B., Voorhees, E.M.: The TREC-5 confusion track: comparing retrieval methods for scanned text. Inf. Retrieval **2**(2), 165–176 (2000). https://doi.org/10.1023/A:1009902609570
8. Needleman, S., Wunsch, C.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48**, 443–453 (1970)
9. Nguyen, T., Jatowt, A., Coustaty, M., Nguyen, N., Doucet, A.: Deep statistical analysis of OCR errors for effective post-OCR processing. In: ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 29–38, June 2019

10. Parapar, J., Freire, A., Barreiro, Á.: Revisiting N-gram based models for retrieval in degraded large collections. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) ECIR 2009. LNCS, vol. 5478, pp. 680–684. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-00958-7_66
11. Peters, C., Braschler, M.: European research letter: cross-language system evaluation: the CLEF campaigns. J. Am. Soc. Inf. Sci. Technol. **52**(12), 1067–1072 (2001)
12. Rigaud, C., Doucet, A., Coustaty, M., Moreux, J.P.: ICDAR 2019 competition on post-OCR text correction. In: International Conference on Document Analysis and Recognition (ICDAR) (2019)
13. Taghva, K., Borsack, J., Condit, A.: Evaluation of model-based retrieval effectiveness with OCR text. ACM Trans. Inf. Syst. **14**(1), 64–93 (1996)
14. Tanner, S., Muñoz, T., Ros, P.H.: Measuring mass text digitization quality and usefulness: lessons learned from assessing the OCR accuracy of the British library's 19th century online newspaper archive. D-Lib Mag. **15**(7/8), 1082–9873 (2009)