International Journal on Digital Libraries (2023) 24:45–62
https://doi.org/10.1007/s00799-023-00345-6

Evaluating and mitigating the impact of OCR errors on information retrieval

Lucas Lima de Oliveira[1] · Danny Suarez Vargas[1] · Antônio Marcelo Azevedo Alexandre[2,3] · Fábio Corrêa Cordeiro[2,4] · Diogo da Silva Magalhães Gomes[2] · Max de Castro Rodrigues[2] · Regis Kruel Romeu[2] · Viviane Pereira Moreira[1]

Received: 5 July 2022 / Revised: 3 January 2023 / Accepted: 7 January 2023 / Published online: 26 January 2023

Abstract
Optical character recognition (OCR) is typically used to extract the textual contents of scanned texts. The output of OCR can be noisy, especially when the quality of the scanned image is poor, which in turn can impact downstream tasks such as information retrieval (IR). Post-processing OCR-ed documents is an alternative to fix digitization errors and, intuitively, improve the results of downstream tasks. This work evaluates the impact of OCR digitization and correction on IR. We compared different digitization and correction methods on real OCR-ed data from an IR test collection with 22k documents and 34 query topics on the geoscientific domain in Portuguese. Our results have shown significant differences in IR metrics for the different digitization methods (up to 5 percentage points in terms of mean average precision). Regarding the impact of error correction, our results showed that on the average for the complete set of query topics, retrieval quality metrics change very little. However, a more detailed analysis revealed it improved 19 out of 34 query topics. Our findings indicate that, contrary to previous work, long documents are impacted by OCR errors.

Keywords Information retrieval · OCR errors · Error correction · Geoscientific documents

1 Introduction

A significant part of textual information exchanged in digital documents, such as scientific articles, technical reports, project proposals, and contracts, is typically stored and distributed in portable document format (PDF). Searches for documents in the PDF format are increasingly common.[1] A report from the PDF association[2] has some impressive statistics that confirm the wide adoption of this format. In 2016, there were over 2 billion PDF documents on the public Web and over 20 billion in Dropbox. About 60% of non-image

✉ Viviane Pereira Moreira
viviane@inf.ufrgs.br

Lucas Lima de Oliveira
lloliveira@inf.ufrgs.br

Danny Suarez Vargas
dsvargas@inf.ufrgs.br

Antônio Marcelo Azevedo Alexandre
am.azevedo@petrobras.com.br

Fábio Corrêa Cordeiro
fabio.cordeiro@petrobras.com.br

Diogo da Silva Magalhães Gomes
diogo.gomes@petrobras.com.br

Max de Castro Rodrigues
max.rodrigues@petrobras.com.br

Regis Kruel Romeu
regisromeu@gmail.com

[1] Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

[2] Petrobras Research and Development Center (CENPES), Rio de Janeiro, RJ, Brazil

[3] Systems Engineering and Computer Science Program (PESC/COPPE), Federal University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil

[4] School of Applied Mathematics, Getulio Vargas Foundation, Rio de Janeiro, RJ, Brazil

[1] According to GoogleTrends https://trends.google.com/trends/explore?date=all&q=%2Fm%2F0600q.

[2] https://www.pdfa.org/wp-content/uploads/2018/06/1330_Johnson.pdf.

files sent as e-mail attachments in Outlook Exchange Enterprise were in PDF.

Before being fed to natural language processing (NLP) or information retrieval (IR) algorithms, the textual contents of these files need to be extracted. When the PDF file was not digitally created (*i.e.,* if it was scanned), the digitization process involves the use of optical character recognition (OCR) algorithms to identify the textual elements within the image.

Industries with long project life cycles, such as Oil & Energy and Mining, rely on old reports that were not digitally created and still are an important target for IR applications. Geological analysis may rely on ancient reports as some oilfields continue to be productive more than 80 years after their discovery. [7], which shares some co-authors with our work, analyzed a sample of 55K documents from the digital library of a petroleum company and found documents that were processed with a wide range of OCR technologies over the years. The quality of these documents varies considerably, and 20% of them were scanned but did not include any embedded text. Moreover, because digital transformation maturity is uneven across different countries and industry sectors, scanned documents are still produced today on a daily basis in different proportions, albeit in ever smaller numbers. In corporate offices, most commercially available multifunction (all-in-one) printers with scanning capabilities and automatic document feeders conveniently produce digital copies of documents in PDF, and automatically send them to the user at the end of the scanning process. By examining the metadata from that very same sample of scanned PDF documents, we found that at least 4% were generated by such scanners. Many of those were rubber-stamped and contained handwritten signatures, which may explain why the scanned documents were preferred over the original digital files for archival purposes.

Although OCR technology has been improving over the years, it is still not perfect. Furthermore, the quality of scanned text may be poor, especially for older documents. [48] estimated that, with an accuracy rate of 99% at the character level (assuming an average word length of five characters), one in 20 words would have a digitization error (*i.e.,* a 5% word error rate).[3] [1] demonstrated that starting at a 5% word error rate, significant impacts are noticed in retrieval quality.

Figure 1 shows an excerpt of real OCR digitization errors in a document from the REGIS collection [44], an IR collection on the geoscientific domain. The original PDF document, Fig. 1a, was processed by Apache Tika to extract its textual contents, which are shown in Fig. 1b. Digitization errors are highlighted in orange, and their counterparts in the original PDF are in green. All occurrences of *reservatório* (reservoir) were erroneously extracted to *reservat6rio*. As a result, queries with the keyword *reservatório* would not be able to retrieve this document. While this type of error would be easier to detect (since it generated an invalid word, *i.e.,* which does not exist in the Portuguese vocabulary), some errors end up generating valid words and are harder to identify. This is the case of *clásticos* (clastic),[4] in line 1, incorrectly extracted as *elásticos* (elastic), which is a valid word in Portuguese with a completely different meaning. [1] showed that this type of error can be found even in mainstream search engines such as Google Scholar. These issues have been motivating a new wave of recent approaches for post-OCR text correction [11,20,37,55].

Although the impacts of OCR-ed text in IR have already been studied [1,10,17,28,53], there is not much work on evaluating the impact of post-processing techniques that try to fix digitization errors. Our earlier work [55] showed that spelling correction was able to improve retrieval results in a news collection. However, the experiments relied on a dataset containing synthetically inserted errors aiming to mimic the most common error patterns found in real systems.

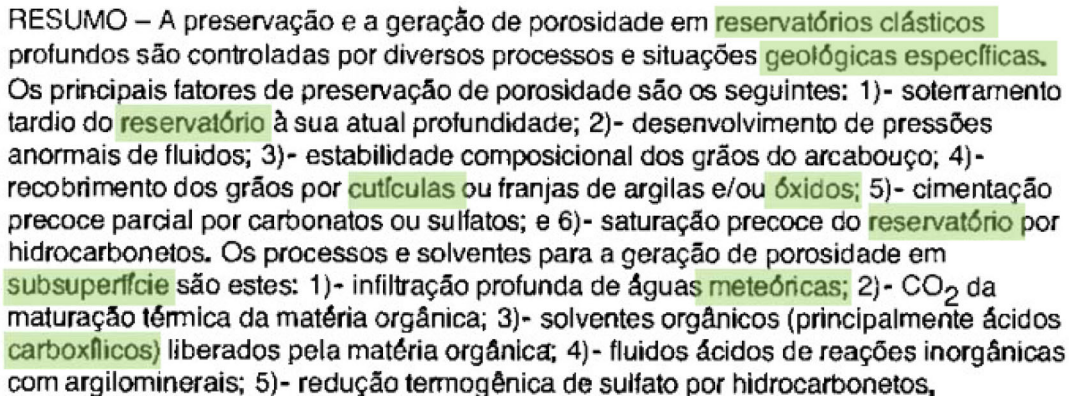
Another important issue concerns the language used in the experiments. As expected, the vast majority of the works were done over English texts. English is a simple language with 26 letters and no diacritics and OCR is typically language-dependent. Therefore, in order to increase representativity, it is crucial to conduct further research to build resources and develop solutions that can address languages other than English [2]. A recent survey on this topic [42] concludes by suggesting that upcoming work on this topic should focus on post-OCR processing in other languages. Fortunately, new datasets in other languages are being generated. In the scope of the ICDAR 2019 competition on post-OCR text correction [45], datasets in 10 European languages (Bulgarian, Czech, Dutch, English, Finish, French, German, Polish, Spanish, and Slovak) were made available. In this article, our target language is Portuguese, which despite being the 6th language in the number of native speakers, does not count with a public dataset for evaluating real OCR errors.

In recent years, NLP in the geoscientific domain has been gaining attention [9,18,31]. However, works addressing IR in this specific domain are rare. This article addresses this gap by experimenting with IR on a geoscientific document collection.

Our research goal in this article is to answer two main questions: (*i*) How does the quality of the digitization affect retrieval results? and (*ii*) Can post-processing OCR-ed texts improve retrieval quality? We performed experiments with

[3] A digitization error happens when the OCR software fails to correctly recognize the characters in the input document. This is different from misspellings, which are human-generated.

[4] Clastic is an adjective that describes a type of rock consisting of broken pieces of other rocks (Cambridge Dictionary).



(a) Original PDF Document

RESUMO - A preservação e a geração de porosidade em reservat6rios elásticos profundos são controladas por diversos processos e situações geol6gicas especflicas. Os principais fatores de preservação de porosidade são os seguintes: 1)- soterramento tardio do reservat6rio à sua atual profundidade; 2)- desenvolvimento de pressões anormais de fluidos; 3)- estabilidade composicional dos grãos do arcabouço; 4)- recobrimento dos grãos por cutrculas ou franjas de argilas e/ou 6xidos; 5)- cimentação precoce parcial por carbonatos ou sulfatos; e 6)- saturação precoce do reservat6rio por hidrocarbonetos. Os processos e solventes para a geração de porosidade em subsuperffcie são estes: 1)- infiltração profunda de águas mete6ricas; 2)- CO2 da maturação térmica da matéria orgânica; 3)- solventes orgânicos (principalmente ácidos carboxnicos) liberados pela matéria orgânica; 4)- fluidos ácidos de reações inorgânicas com argilominerais; 5)- redução termogênica de sulfato por hidrocarbonetos,

(b) Extracted Textual Contents

Fig. 1 Example of OCR errors in a document from the REGIS collection. Extraction errors and their corresponding inputs are highlighted

REGIS [44], an IR test collection composed of PDF documents in Portuguese for the geoscientific domain, along with query topics and their corresponding relevance assessments. Documents are very long, were produced over a long time span (1957 to 2020), and vary substantially in terms of visual quality. We evaluated three digitization tools with OCR capabilities and two error correction methods. Contrary to existing work that argues that long documents are robust to OCR errors [10,39,52], we found that retrieval quality metrics varied significantly depending on the digitization system. For error correction, our results showed that on average for the complete set of query topics, retrieval quality metrics change very little. However, a more detailed analysis showed that most query topics (19 out of 34) improved with error correction.

The contributions of this article are:

- An investigation of the impact of different digitization and correction methods for OCR-ed texts using real OCR-ed data.
- An evaluation of the intrinsic quality of text digitization and error correction.
- Experiments with a language that, despite being widely spoken, is underrepresented in terms of IR resources.
- Experiments on a domain that remains largely underexplored in IR research.

The remainder of this article is organized as follows. Section 2 discusses background and related work. Section 3 details the methodology adopted in our investigation. Section 4 reports the results of our experiments. Finally, Sect. 5 concludes this article and discusses future directions.

2 Background and related work

OCR is the process of automatically extracting text present in digital images. Although OCR methods have been studied for a long time, they are still imperfect, especially if the input documents were captured with poor image quality.

There are two types of word errors generated by the OCR process – *non-words*, when the extracted word does not exist, and *real-words*, when the incorrectly extracted word corresponds to a valid word. [41] analyzed English and French texts and observed that 60% of the errors are from *real-words*, while 40% are from *non-words*. Additionally, some of these errors can also be classified as *segmentation errors*. An *incorrect segmentation* happens when a word is unduly split into two or more, and an *incorrect concatenation* occurs when a space character is suppressed. [41] state that incorrect segmentation is 2.3 times more frequent than incorrect concatenation. Also, these types of errors are not exclusive, both types of segmentation errors can generate *non-words* or *real-words*.

The impact of OCR errors has been studied on a variety of tasks including contextual embeddings [25], named entity recognition [12,13,19,21,22,24,38], entity linking [33,34], part-of-speech tagging [32], text summarization [26], text classification [22,58,59], topic modeling [22,40], and event detection [4]. The common finding is that digitization quality impacts the results of the downstream tasks and that mechanisms to mitigate errors are successful.

Specifically for IR, the pioneer studies that aimed at assessing the impact of OCR-ed text date back to the 1990s and early 2000s [10,39,51–53].

Taghva et al. devoted a significant amount of work into this topic. Initially, in [51], the authors report on experiments using a collection with 204 documents and 71 query topics. Both OCR-ed version and ground truth texts were indexed in a Boolean retrieval system. No relevance judgments were available, so the comparison was between retrieval using ground truth and OCR-ed documents. The results showed a 97.6% overlap. The main finding was that retrieval was robust to OCR errors especially because the documents were long (38 pages on average) and thus were likely to have a correct version of the queried term. They designed a simple correction tool based on syntactic similarity, which enabled an increase of one percentage point in terms of retrieved documents. Nevertheless, because the retrieval system did not produce ranked results, the impact of OCR errors and their correction could not be gauged. Subsequently, in [52], the authors generated relevance judgments for the query topics and used them in a ranked retrieval setting (*i.e.,* the vector space model). The conclusion was that average precision was not affected by OCR errors and that post-processing did not bring significant improvements. Finally, in [53], the authors worked with full-text and abstract-length documents and with simulated and actual OCR errors. The finding was that as document length decreases, the errors have a higher impact on precision. However, the impact was only significant in a collection with very small documents (*i.e.,* with an average of 50 words per document).

[10] after experimenting with simulated OCR errors on four IR collections, also found that small documents are more affected by OCR errors. While the collection with the longest documents had a 4% decrease in average precision, in the collection with the shortest documents, the decrease was 10%.

[39] conducted a probabilistic analysis on the impact of errors on IR quality. They took a theoretical perspective and modeled OCR errors as a random process. Their conclusion was that IR is robust to many errors and that spelling correction based on dictionaries should not be performed as it would not improve retrieval results.

In summary, the consensus drawn from these early experiments was that retrieval was robust to OCR errors, especially for long documents. However, there are also studies with findings that go in the opposite direction. [56] experimented with post-processing and found an improvement of 8.5 percentage points in word recognition. However, their experiments focused on indexing (and not on retrieval) so there were no queries involved. Similarly, [14] worked with historical texts in three versions (extracted, ground truth, and automatically corrected). The authors found an improvement of almost 60% in recall misses (*i.e.,* the number of unique ground-truth words in the automatically corrected version was 60% higher than the number found in the uncorrected version). Along the same lines, [54] studied the impacts of correcting OCR errors on document retrieval. They found that high error rates correlate with low retrievability scores (*i.e.,* a metric of how often a document occurs when inspecting the top-k results for a set of queries) and that error correction leads to higher retrievability. In a similar fashion, [50] looked at score changes in the ranking. They observed that, as expected, the divergence in relation to the ranking generated for the ground-truth documents increases as OCR quality decreases.

Thus far, most of the existing work on assessing the impact of OCR-ed text and correction methods in IR can be divided into two groups—(i) works using IR test collections (*i.e.,* with documents, query topics, and relevance judgments), which relied on *synthetically created* OCR errors [1,10]; and (ii) works using real OCR-ed documents, which lacked query topics and/or relevance judgments [14,50,51,54,56]. To the best of our knowledge, only five works experimented with real OCR errors in a true IR setting, namely the pioneer works by [52,53] which were previously described in this section, and the works by [17,30], and [29].

[30] took an interesting approach to assemble their data collection—they took the TREC-8 spoken document retrieval collection [27], generated manual transcriptions of the audio data, printed them, scanned the printed documents,

and then processed them with an OCR software. Their experiments revealed that query expansion was more affected by the OCR errors than the baseline retrieval run.

[17] worked with documents from the FIRE RISOT collection in Bangla (68 K documents and 66 query topics) and Hindi (94 k documents and 28 query topics). They found that the difference in average precision between the extracted and ground truth versions was very large (31% for Bangla and 57% for Hindi). They tested several mechanisms to expand the query with the goal of improving retrieval performance and, while these approaches were successful, they were still far from the results achieved on the ground-truth documents.

Recent work by [29] examined the impact of OCR quality in an interactive IR scenario. The search was done over historic news articles in Finish that were digitized by two different systems. The authors observed that the users assigned higher relevance scores to documents with improved digitization.

The improvements found by [17] and the lack of difference found by other research on English data [10,30,51–53], may suggest that language impacts the results. A recent survey by [42] reports on 17 openly accessible datasets with OCR-ed texts and their ground truths. There are 13 languages covered by these datasets: English, Dutch, French, Latin, Bulgarian, Czech, Finish, German, Polish, Spanish, Slovak, Italian, and Romansh. These datasets are being used to aid in the development of several post-OCR correction approaches based on a large set of underlying techniques—from lexical approaches to language models. However, the language we use in this article, Portuguese, is not covered by existing OCR datasets.

Table 1 presents an overview of all works that investigated the impact of OCR in IR that we could find in our literature review. In the last row, we show how our work fits into the analyzed features. We assessed seven characteristics of these works, namely (i) whether each publication dealt with real OCR errors (*i.e.,* if they used real PDF documents); (ii) if they had the corresponding ground truth (*i.e.,* text without errors); (iii) if they followed the standard IR evaluation procedure with queries and relevance judgments; (iv) the length of the documents used (where "L" corresponds to long, and "S" to short); (v) whether they addressed OCR error correction; (vi) which languages were used in the experiments; (vii) and the domain of the test collection used in the experiments. It can be seen that half the works that conducted their experiments with real PDF documents did not use queries and relevance judgments [14,29,50,51,54,56]. On the other hand, most works that followed the standard IR experimental procedure [1,10,28,39,55,57] relied on synthetic OCR errors. Reinforcing the considerations made by [42] and [2], the majority of these works (10 out of 16) used only collections in English. Regarding document length, most works (10 out of 16) used only short documents in their experiments. As for the domain, news articles are by far the most widely used type of document (8 out of 16).

More technical domains are also present in recent investigations of digitization. Engineering documents were used by [16]. The authors performed text detection and post-OCR correction on industrial maps. They found that post-OCR correction led to an improvement of 7 percentage points in text recognition. This work, unlike ours, does not deal with IR.

In this article, we build upon our previous work [1,55]. We move to a realistic scenario by working with real digitization errors in a different IR collection. In comparison with the existing work presented in Table 1, our evaluation has the complete set of elements that enable assessing the impact of OCR digitization and error correction in a traditional IR experimental setting.

3 Materials and methods

To assess the impact of OCR digitization and post-processing on IR quality, we follow the process depicted in Fig. 2. We start by taking the original PDF documents and extracting their textual contents (1). Once the text is extracted, it is submitted to a post-processing step that aims at fixing digitization errors (2). Then, at the Evaluation stage (3), the queries are run and scored using the relevance judgments.

The pipeline described in Fig. 2 can be seen as an *extrinsic evaluation* since it assesses how digitization/correction impacts a downstream task. Nevertheless, it is also important to have an *intrinsic* evaluation to provide insights into the inherent quality of the digitization and correction processes, regardless of the retrieval results. The intrinsic evaluation pipeline is described in Fig. 3, and it requires a ground truth. The main phases for these experiments are similar to the ones in Fig. 2. The difference is in the evaluation step, which uses the ground truth digitization instead of queries and relevance judgments. In the next subsections, we describe in detail each of these phases, as well as the materials and methods used in our experiments.

3.1 Test collection

The ideal test collection for our experimental setting would require four components to enable a complete evaluation of the quality of the digitization and its impact on retrieval: (i) real PDF documents, (ii) the expected textual output *i.e.,* the ground truth, (iii) query topics, and (iv) relevance judgments. Existing test collections created within the scope of post-OCR competitions such as ICDAR [8,45] are not suitable because they lack query topics and relevance judgments (*i.e.,* components iii and iv). IR test collections typically

Table 1 Comparison of related works that evaluate the impact of OCR errors for information retrieval

Work	Real PDFs	Ground truth	Doc. Length	Relevance judgments	Error correction	Language	Domain
[10]	×	✓	S, L	✓	×	EN	Energy
[51]	✓	✓	S, L	×	✓	EN	Nuclear Regulations
[56]	✓	✓	S	×	✓	DE	Technical Abstracts
[52]	✓	✓	L	✓	✓	EN	Energy
[53]	✓	✓	S, L	✓	✓	EN	Energy, Computer Science, Law, News
[28]	×	✓	S	✓	✓	EN	Governmental Docs
[39]	×	✓	S	✓	✓	EN	Governmental Docs
[30]	✓	✓	S	✓	×	EN	News
[14]	✓	✓	S	×	✓	EN	News
[17]	✓	✓*	S, L	✓	×	EN, BN, HI	News, Legal Docs, Governmental Docs
[54]	✓	✓	S	×	✓	NL	News
[50]	✓	✓	S	×	×	EN	News
[1]	×	✓	S	✓	×	PT	News
[55]	×	✓	S	✓	✓	PT	News
[57]	×	✓	S	✓	×	EN	Passages from Web documents
[29]	✓	×	S	✓	✓	FI	News
This work	✓	✓*	S, L	✓	✓	PT	Geosciences

An asterisk (*) indicates that the ground truth does not cover the entire set of documents

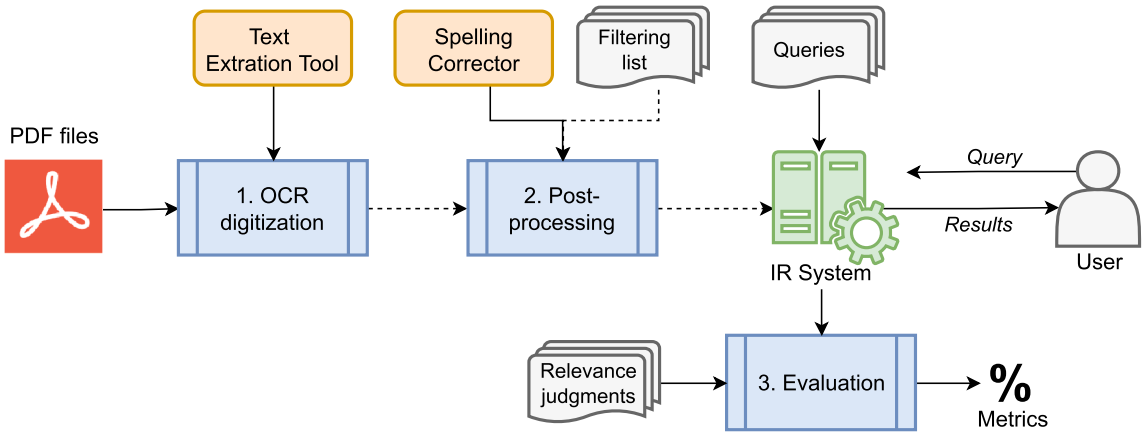


Fig. 2 Pipeline for the extrinsic evaluation of the impact of text digitization and correction on information retrieval

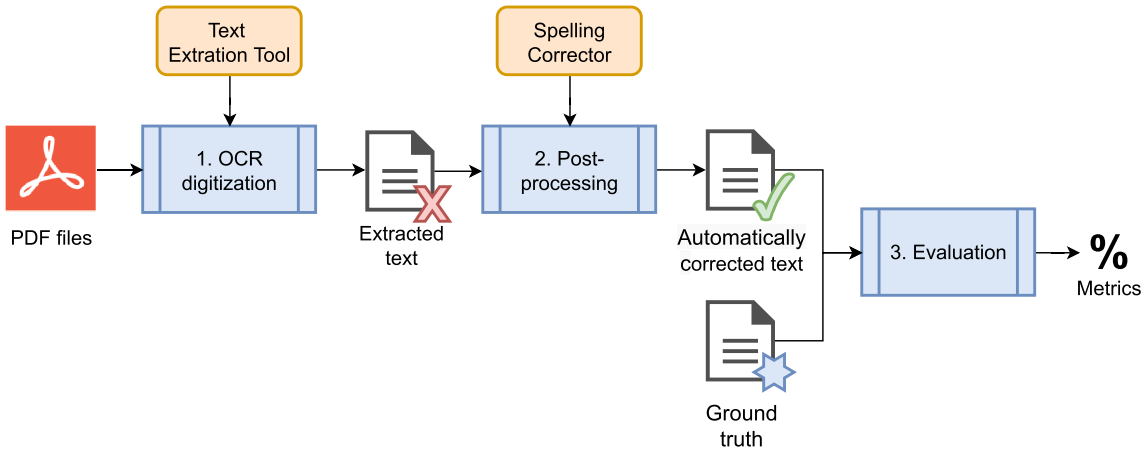


Fig. 3 Pipeline for the intrinsic evaluation methodology

lack component *i*, as they work with text that was originally in a digital format.

The test collection used in our experiments is the recently released **REGIS** (Retrieval Evaluation for Geoscientific Information Systems) collection [44].[5] REGIS is composed of 21K geoscientific documents in Portuguese stored in PDF files. Samples of pages of these documents are in Fig. 4. The files were created over a long period of time, between 1957 and 2020. A portion of the files consists of scanned documents and requires OCR. Analyzing the metadata of the documents (including author, creation, and production applications) we concluded that 38.44% of the collection (*i.e.,* 8244 documents) were scanned and the remaining 61.56% are born-digital. When we consider only the documents that were judged as relevant for the queries, we found an even spread between scanned (46%) and born-digital (54%). The documents consist of technical reports, theses, and dissertations—which means they are typically very long, with an average of 25.1k tokens per document (111 pages). A histogram of the number of pages in our documents showed a clear bimodality. The first population, classified here as short, is formed mainly by articles. These, in general, have up to 30 pages. The second population, classified here as long documents, is composed of theses and dissertations. Thus, setting the threshold to 30 pages, most of our documents (78%) were considered long. When we look at the distribution of long and short among the relevant documents, we found that 72% of those are longer than 30 pages. The complete statistics are in Table 2. Finally, another important characteristic of the documents is that they are filled with technical terms and there are many figures, tables, chemical formulas, and equations. These issues are likely to pose an added difficulty for retrieval systems.

REGIS lacks the ground truth digitizations (component *ii*) that are required by the intrinsic evaluation (Fig. 3). It would be extremely costly to manually generate ground truth digitizations for the entire collection (over 2.4 million PDF pages) since this involves laborious manual checking. As a result, we created ground truth digitizations for a sample of sentences selected from different relevant documents cover-

[5] https://github.com/Petroles/regis-collection.

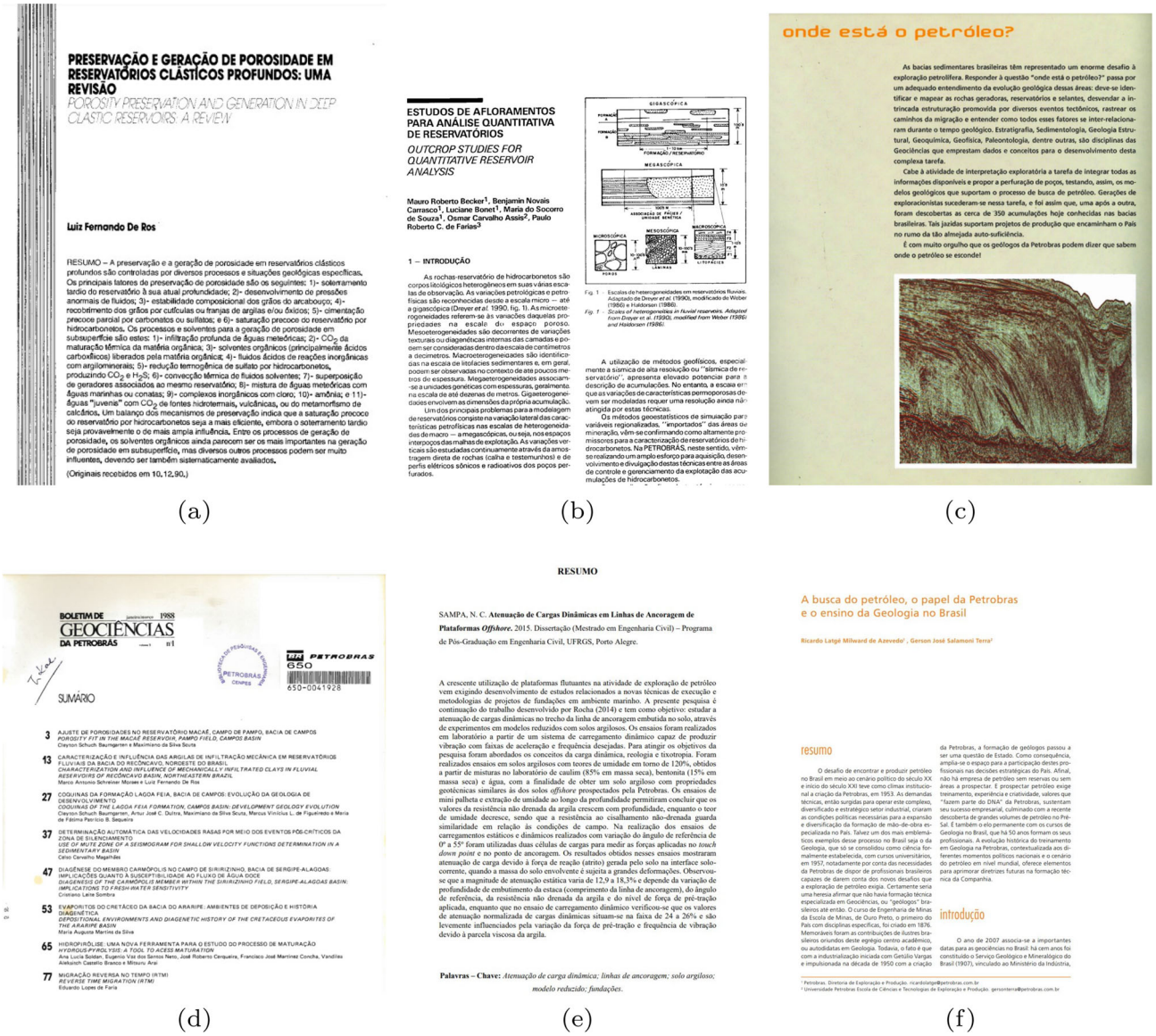


Fig. 4 Example of documents in REGIS. The images in **a**, **b**, **c**, and **d** correspond to scanned documents, while **e** and **f** are born-digital

ing all query topics. The rationale for selecting the sample among relevant documents was to focus on documents in which digitization errors would have a greater impact. We randomly chose sentences that contained at least one of the keywords in the query topic. These sentences were checked by two human annotators. Our ground truth covers 170 sentences with an average of 30 tokens by sentence, and it is publicly available.[6]

3.2 Digitization tools

All documents in REGIS were submitted to a set of digitization tools that are responsible for performing step 1 in our methodology (Figs. 2 and 3). In our choice of tools, we aimed at ensuring some variety. Thus, we included both open-source and commercial software. Also, we chose among tools that can readily work for Portuguese texts without the need for further training. Recall that 61% of REGIS consists of born-digital documents. The decision regarding whether OCR is needed is made by the digitization tools themselves. The tools used in our experiments were:

- **Apache Tika**,[7] an open-source application that can be used through the command line to parse and extract information from many different file types. It uses the

[6] https://github.com/lucaslioli/regis-collection-gs.

[7] https://tika.apache.org/.

Table 2 Statistics of the REGIS IR collection

Documents	21,444
Tokens in documents	538.4 M
Avg tokens per document	25.1 K
Distinct tokens	4.1 M
Avg pages per document	111
Total pages	2.4 M
Born-digital documents	13,200 (61.56%)
Scanned documents	8244 (38.44%)
Short documents	4716 (21.66%)
Long documents	16,728 (78.34%)
Born-digital Short documents	2462 (11.48%)
Born-digital Long documents	10,738 (50.07%)
Scanned Short documents	2183 (10.18%)
Scanned Long documents	6061 (28.26%)
Query topics	34
Avg tokens per topic (title)	5.53
Avg tokens per topic (description)	31.32
Avg tokens per topic (narrative)	76.85

Tesseract OCR Parser.[8] Tesseract is a popular solution that was extensively used in several of the related works [21,22,33,36]. Tika has a PDF parser that automatically chooses if the document needs OCR by first trying to extract the textual contents in a page. If fewer than ten characters are obtained or if there are more than ten characters with unmapped Unicode values, OCR is used.

- **ABBYY FineReader**,[9] a desktop proprietary software with many features to manipulate and process PDF files. We could not find the description of how this tool decides whether OCR is necessary. ABBYY produces state-of-the-art results and has been used in related work [29,36] and in nine works surveyed by [42].
- **Tornado**,[10] a tool developed by the Brazilian oil company (Petrobras) and PUC-Rio. Tornado is developed in Python and built upon the following tools and libraries: Poppler,[11] Detectron2,[12] PDFMiner,[13] Camelot,[14] Tesseract OCR[8], and Luigi.[15] It relies on machine learning to selectively extract information from PDF files: not only text, but also figures, charts, and tables. This choice was motivated by the fact that Tornado is tailored for document digitization in the Oil and Gas industry, which includes the geoscientific domain of the REGIS collection. The tool is able to perform visual grouping of elements on a PDF page (*i.e.,* blocks of text, figures, charts, or tables) by using computer vision detection and segmentation algorithms. When a block of text is detected, native text extraction from the page is attempted first by filtering the characters whose coordinates are in the range of the detected bounding box. If the resulting filtered text is an empty string or if it contains a number of Non-Unicode glyph mapping indices above a given threshold, the image corresponding to the text block area is submitted to the image-enhancing pipeline described by [7], and then OCR. Finally, texts from all blocks are combined into a single sequence of characters.

3.3 Post-processing methods

Two methods were used to post-process the OCR-ed texts aiming to fix digitization errors, namely:

- SymSpell,[16] a language-independent spelling corrector that uses the Symmetric Delete algorithm with the aim to achieve faster processing, based on predefined dictionary lookups of unigrams and bigrams and Levenshtein edit distance.
- sOCRates [55] is a recently released post-OCR text corrector developed using Portuguese texts. It relies on a BERT-based classifier trained to identify sentences with errors and on a second classifier that relies on format, semantic, and syntactic similarity features.

Post-processing can be time-consuming. Thus, aiming to improve efficiency, a filtering list was used to avoid post-processing the born-digital PDF documents. This way, only the scanned documents in REGIS went through the correction step. As reported in Sect. 3.1, these documents correspond to 38% of the collection and 46% of the relevant documents.

3.4 IR system

The IR system used in our experiments was Apache Solr,[17] which is an open-source search platform based on Apache Lucene. The configurations adopted were the same as the best result obtained to create REGIS [44], with BM25 as the scoring function and Portuguese Light Stemmer. Despite having been proposed over 20 years ago, BM25 is still a widely used baseline for modern methods, including important work on

[8] https://github.com/tesseract-ocr/tesseract.

[9] https://www.abbyy.com/.

[10] https://petroles.puc-rio.ai/index_en.html, see tab *Development in progress*.

[11] https://github.com/freedesktop/poppler.

[12] https://github.com/facebookresearch/detectron2.

[13] https://github.com/pdfminer/pdfminer.six.

[14] https://github.com/camelot-dev/camelot.

[15] https://github.com/spotify/luigi.

[16] https://github.com/wolfgarbe/SymSpell.

[17] https://lucene.apache.org/solr/.

neural reranking [43] and recent techniques for document expansion [35]. Proximity search was used in the retrieval phase to give higher scores to documents in which the query terms appear in close proximity—this is important since our documents are typically very long. Queries consisted of the contents of the title field of the topics, and 100 documents were retrieved for each query.

3.5 Evaluation metrics

IR evaluation metrics were computed using Trec_eval.[18] Our analysis focused on *Mean Average Precision* (MAP) and Normalized Discounted Cumulative Gain (*NDCG*) which are the standard metrics to evaluate ranked results. In addition, it is important to consider a recall-oriented metric to assess how many relevant documents failed to be retrieved in the different experimental runs. Thus, like [30], we used the number of relevant documents that were retrieved (*Rel. Ret*). To assess whether the differences were statistically significant, we used t-tests with $\alpha = 0.05$. T-tests have been widely used in IR to compare means [1,15,57] and their appropriateness has been confirmed by different studies [23,46,49]. A Shapiro-Wilk normality test verified that our MAP scores follow a normal distribution. We did not include precision at different cut-off values in our analyses because these metrics are less stable, hence they require more query topics to yield reliable scores [5].

For the intrinsic experiments, we calculated *Character Error Rate* (CER) and *Word Error Rate* (WER) [6] using the OCREvaluation script.[19] These metrics are widely used to evaluate OCR post-processing methods. They take into account the number of differences at character and word level between the output and the ground truth version.

Relevance assessments in REGIS are graded on four levels: "very relevant", "fairly relevant", "marginally relevant", and "not relevant". We experimented with two scenarios to calculate the retrieval metrics that rely on binary assessments—a *tolerant* scenario in which marginally relevant documents are considered relevant and a *strict* scenario in which documents need to be at least fairly relevant to be classified as relevant.[20]

[18] https://trec.nist.gov/trec_eval/.

[19] https://github.com/impactcentre/ocrevalUAtion.

[20] The strict and tolerant scenarios only affect the metrics that use binary relevance judgments (*i.e.,* relevant/not relevant. MAP is one of such metrics. NDCG, on the other hand, works by definition with multiple levels of relevance.

Table 3 Information retrieval quality metrics for the OCR digitization systems

Configuration	MAP	NDCG	Rel.Ret
Tolerant scenario			
(1) Tika	.4947	.6705	657
(2) ABBYY	**.5438**	**.7109**	**697**
(3) Tornado	.5054	.6911	666
Strict scenario			
(4) Tika	.4549	.6705	420
(5) ABBYY	**.4901**	**.7109**	**442**
(6) Tornado	.4636	.6911	427

The tolerant scenario considers "Marginally Relevant" and the strict scenario considers "Fairly Relevant" as the minimum relevance level. Best results in bold

4 Experimental results

In this section, we present the results of our experimental evaluation analyzed under different perspectives to answer our research questions.

4.1 The impact of OCR quality on retrieval effectiveness metrics

Table 3 shows the results for three digitization systems (Tika, ABBYY, and Tornado) across both scenarios. As expected, the scores in the *strict* scenario are lower since documents need to be graded at least as fairly relevant to be classified as relevant by the metrics that rely on binary judgments. The best digitization system was clearly ABBYY. It was the best performing according to all metrics in both scenarios. Tika and Tornado had similar MAP scores, with Tornado being able to retrieve more relevant documents and achieving a higher NDCG. ABBYY's superior scores were statistically significant for MAP, Rel. Ret., and NDCG. No statistically significant differences were found between Tika and Tornado for any of the retrieval metrics. As a disadvantage, ABBYY is a commercial software, and the version we acquired has limitations in terms of the number of pages it could process in a month. Tika is freely available and could easily be integrated into our code, while Tornado is under development and currently not publicly available.

The differences in retrieval scores show that digitization quality indeed impacts retrieval evaluation metrics—even for long documents such as the ones in REGIS.

4.2 Impact of error correction on retrieval results

To analyze the impact of error correction on retrieval performance, we took the outputs of Tika and ABBYY and post-processed them with sOCRates and SymSpell. Tika was

Table 4 Results for OCR error correction

Configuration	MAP	Δ	NDCG	Δ	Rel.Ret	Δ
Tolerant scenario						
(1) Tika	**.4947**	–	**.6705**	–	**657**	–
(2) + sOCRates	.4904	−0.87%	.6664	−0.61%	652	−0.76%
(3) + SymSpell	.4810	−2.77%	.6566	−2.07%	648	−1.37%
(4) ABBYY	**.5438**	–	**.7109**	–	**697**	–
(5) + sOCRates	.5438	0.00%	.7109	0.00%	697	0.00%
(6) + SymSpell	.5173	−4.87%	.6869	−3.38%	688	−1.29%
Strict scenario						
(7) Tika	**.4549**	–	**.6705**	–	**420**	–
(8) + sOCRates	.4428	−2.66%	.6664	−0.61%	417	−0.71%
(9) + SymSpell	.4313	−5.19%	.6566	−2.07%	417	−0.71%
(10) ABBYY	**.4901**	–	**.7109**	–	**442**	–
(11) + sOCRates	.4901	0.00%	.7109	0.00%	442	0.00%
(12) + SymSpell	.4483	−8.53%	.6869	−3.38%	437	−1.13%

The tolerant scenario considers "Marginally Relevant" and the strict scenario considers "Fairly Relevant" as the minimum relevance level

chosen because it had the lowest scores in Table 3, which means it has more room for improvement. ABBYY, on the other hand, had the best results, so it allows us to see if post-processing is able to improve results that are already good.

The results of this analysis are in Table 4. Note that the scenarios do not affect the results for NDCG as it does not rely on binary relevance judgments.

In the tolerant scenario, sOCRates was better than SymSpell in all three metrics. In the strict scenario, only in terms of MAP and NDCG. On the other hand, considering efficiency, SymSpell works approximately nine times faster than sOCRates. Additionally, we can see that sOCRates does not alter the results of ABBYY (comparing lines 4–5 and 10–11). This happens because the error detection step in sOCRates considered nearly all sentences as correct and did not proceed with the correction step. When we look at the rankings generated for the queries, we can see differences in BM25 scores. However, the ordering of the documents remained the same, and as a result, the scores for the evaluation metrics are identical.

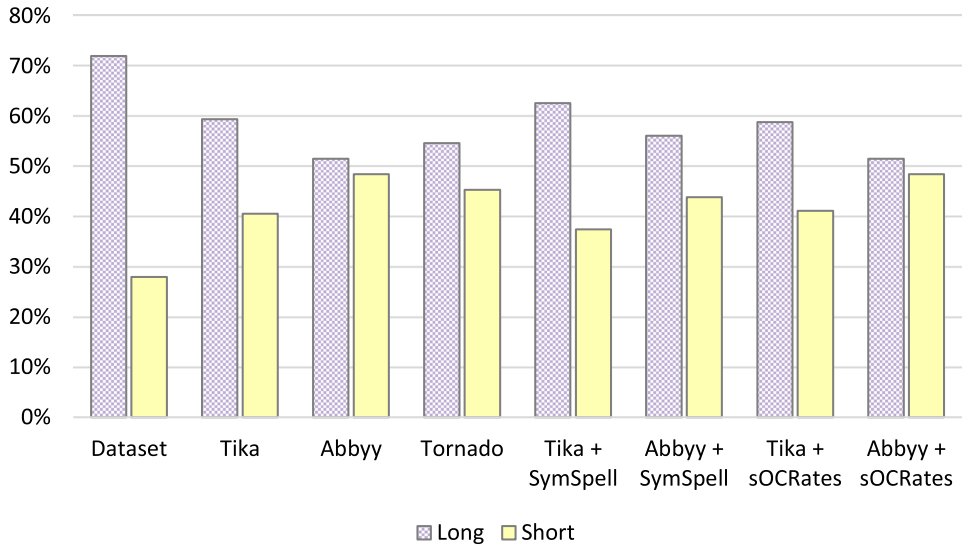
For both scenarios, the correction tools resulted in worse performance compared to the uncorrected runs. However, the differences were not statistically significant. The negative impact of the correction tools was more severe in terms of MAP in the strict scenario. In order to have a better understanding of the reasons behind this, in Sect. 4.3, we perform a topic-by-topic analysis.

It is also interesting to analyze the relevant documents that were not retrieved to see whether they are short or long and whether they were scanned or born-digital. Figure 5 shows the results of this analysis. The first set of bars in both charts reflects the distribution of the type of document within the set of relevant documents. One would expect to find a similar distribution among the documents that failed to be retrieved by the experimental runs. However, when we look at Fig. 5-a, we can see that short documents make up 28% of the relevant set but the rate to which they were unretrieved is always higher (between 37 and 48%). Similarly, scanned documents correspond to 46% of the relevant set, but their prevalence among the documents that failed to be retrieved is always higher (between 50 and 64%). From this analysis, we can conclude that, as expected, short and scanned documents have a greater chance of not being retrieved.

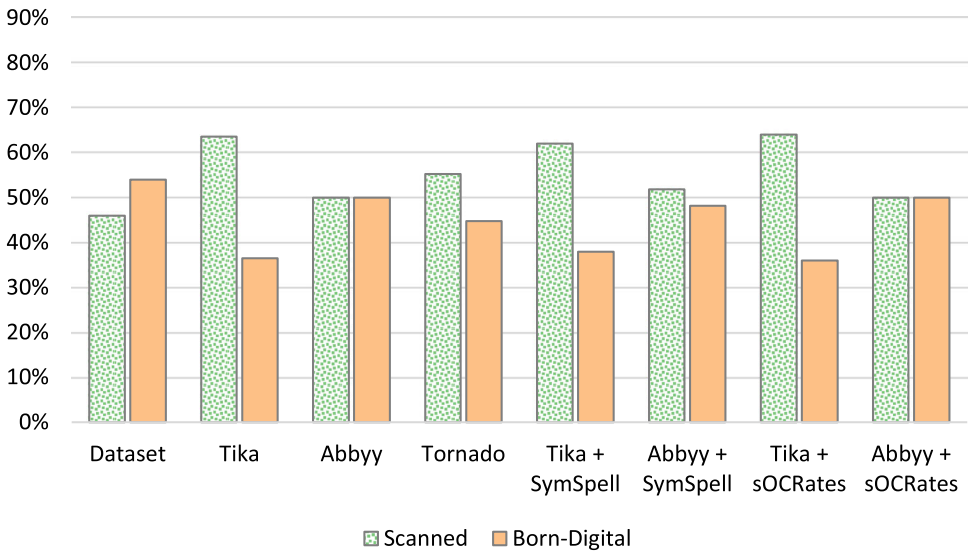
4.3 Topic-by-topic analysis

Figure 6 presents MAP results by topic for the tolerant scenario. Each MAP score is followed by its rank in the seven experimental runs. Top ranks are in darker shades. Our analyses focused on MAP since it is a stable metric for the number of topics we have in the collection [5]. As stated in [44], REGIS has a mix between easier and harder query topics. Queries such as Q7, Q9, Q13, Q22, and Q25 reached higher MAP scores, while queries Q1, Q4, Q16, Q17, and Q31 yielded low scores. This result is also related to the distribution of relevant documents across the query topics, *i.e.,* queries with more relevant documents, tend to have higher scores. As a general tendency, the highest MAP scores were achieved by ABBYY and ABBYY + sOCRates, which can be seen by the large concentration of darker cells in these columns.

Looking at the absolute scores in Figure 6, 19 out of 34 query topics were positively influenced by text correction with SymSpell over Tika (Q3–Q5, Q15–Q21, Q23, Q26, Q28–Q34). For ABBYY, the number of topics that improved with SymSpell was slightly smaller (17 topics–Q3, Q15–



(a) Long vs. Short Documents



(b) Born-Digital vs. Scanned Documents

Fig. 5 Distribution of long/short and born-digital/scanned documents among the documents that failed to be retrieved in each experimental run. The first set of bars reflects the distribution of the type of document within the set of relevant documents

Q19, Q22–Q24, Q26, Q28–32). Queries Q6 and Q10 were the ones in which text correction with SymSpell had a very negative impact both over Tika and ABBYY (proportional decreases of 80 and 52%, respectively). Looking into the specific issues behind the losses, we found the same problem in both Q6 *Sala de visualização 3D* (3D visualization room) and Q10 "*ajuste de histórico usando dados de sísmica 4D*" (historical adjustment using 4D seismic data)—SymSpell changed "3D" and "4D" to "D". This causes a large loss in terms of semantics and, as a result, relevant documents dropped three positions on average on the ranking for Q6 and seven positions in the ranking for Q10. In addition, three documents were missed in the corrected version of Q6 and four in Q10. The largest gains when using SymSpell for correction were observed in Q17. In this query, there are only three relevant documents. Both Tika and Tika+SymSpell retrieved only one relevant document, in the fifth and third positions, respectively. The improvement in the ranking was simply due to fixing hyphenation errors.

sOCRates was better than SymSpell on the average of the complete set of query topics (Table 4). However, when examining the results for the individual queries, we see a different picture. While SymSpell was able to improve the results for individual topics over both Tika and ABBYY, sOCRates was only able to improve 6 topics over Tika and none over ABBYY.

Query	Tika	ABBYY	Tornado	Tika + sOCRates	ABBYY + sOCRates	Tika + SymSpell	ABBYY + SymSpell
Q1	0.2487 (1)	0.1936 (5)	0.2061 (4)	0.2417 (3)	0.1936 (5)	0.2428 (2)	0.1924 (7)
Q2	0.4434 (1)	0.4289 (4)	0.4164 (7)	0.4331 (3)	0.4289 (4)	0.4417 (2)	0.4225 (6)
Q3	0.1613 (6)	0.1804 (2)	0.1759 (5)	0.1597 (7)	0.1804 (2)	0.1785 (4)	0.1983 (1)
Q4	0.4549 (5)	0.4775 (1)	0.4414 (7)	0.4524 (6)	0.4775 (1)	0.4641 (4)	0.4712 (3)
Q5	0.5020 (7)	0.5779 (2)	0.5468 (5)	0.5170 (6)	0.5779 (2)	0.6051 (1)	0.5752 (4)
Q6	0.4456 (4)	0.4645 (1)	0.3746 (5)	0.4463 (3)	0.4645 (1)	0.0900 (7)	0.0906 (6)
Q7	0.9612 (1)	0.9592 (3)	0.9325 (7)	0.9612 (1)	0.9592 (3)	0.9478 (6)	0.9575 (5)
Q8	0.7639 (4)	0.7799 (1)	0.7289 (7)	0.7631 (5)	0.7799 (1)	0.7486 (6)	0.7695 (3)
Q9	0.7492 (6)	0.8941 (1)	0.7557 (4)	0.7496 (5)	0.8941 (1)	0.7386 (7)	0.8650 (3)
Q10	0.6908 (4)	0.7482 (1)	0.7332 (3)	0.6868 (5)	0.7482 (1)	0.3285 (6)	0.3279 (7)
Q11	0.3788 (4)	0.3884 (1)	0.3273 (7)	0.3734 (5)	0.3884 (1)	0.3642 (6)	0.3852 (3)
Q12	0.4355 (4)	0.4656 (1)	0.4036 (7)	0.4283 (5)	0.4656 (1)	0.4167 (6)	0.4595 (3)
Q13	0.7577 (5)	0.9006 (2)	0.8483 (4)	0.7577 (5)	0.9006 (2)	0.7448 (7)	0.9035 (1)
Q14	0.7175 (5)	0.8344 (1)	0.7622 (4)	0.7028 (7)	0.8344 (1)	0.7094 (6)	0.8013 (3)
Q15	0.3330 (3)	0.3034 (5)	0.2835 (7)	0.3330 (3)	0.3034 (5)	0.3658 (2)	0.3749 (1)
Q16	0.0829 (5)	0.1104 (2)	0.0815 (7)	0.0829 (5)	0.1104 (2)	0.1036 (4)	0.1152 (1)
Q17	0.0667 (4)	0.0667 (4)	0.1111 (1)	0.0667 (4)	0.0667 (4)	0.1111 (1)	0.1111 (1)
Q18	0.1910 (4)	0.1833 (6)	0.2230 (1)	0.1909 (5)	0.1833 (6)	0.1977 (2)	0.1965 (3)
Q19	0.2567 (2)	0.2482 (6)	0.2493 (5)	0.2547 (3)	0.2482 (6)	0.2721 (1)	0.2508 (4)
Q20	0.5212 (6)	0.6468 (1)	0.5645 (4)	0.4980 (7)	0.6468 (1)	0.5316 (5)	0.6338 (3)
Q21	0.5485 (6)	0.6568 (1)	0.5592 (4)	0.5365 (7)	0.6568 (1)	0.5500 (5)	0.5611 (3)
Q22	0.8307 (5)	0.8588 (2)	0.8526 (4)	0.8304 (6)	0.8588 (2)	0.8192 (7)	0.8592 (1)
Q23	0.4495 (6)	0.4813 (3)	0.4551 (5)	0.4494 (7)	0.4813 (3)	0.4819 (2)	0.4828 (1)
Q24	0.5673 (5)	0.6497 (2)	0.5768 (4)	0.5505 (6)	0.6497 (2)	0.5492 (7)	0.6600 (1)
Q25	0.9158 (6)	0.9459 (1)	0.9215 (4)	0.9167 (5)	0.9459 (1)	0.9140 (7)	0.9455 (3)
Q26	0.6127 (4)	0.6072 (5)	0.6137 (3)	0.5706 (7)	0.6072 (5)	0.6205 (1)	0.6141 (2)
Q27	0.6717 (3)	0.6630 (4)	0.6806 (1)	0.6755 (2)	0.6630 (4)	0.6592 (6)	0.6525 (7)
Q28	0.5628 (7)	0.6526 (2)	0.6396 (4)	0.5631 (6)	0.6526 (2)	0.5834 (5)	0.6596 (1)
Q29	0.5902 (6)	0.6496 (3)	0.7157 (1)	0.5902 (6)	0.6496 (3)	0.6221 (5)	0.6649 (2)
Q30	0.5127 (6)	0.7430 (2)	0.5992 (4)	0.5107 (7)	0.7430 (2)	0.5275 (5)	0.7506 (1)
Q31	0.0846 (6)	0.2344 (3)	0.2517 (1)	0.0846 (6)	0.2344 (3)	0.0870 (5)	0.2382 (2)
Q32	0.7117 (5)	0.7580 (2)	0.5160 (7)	0.7018 (6)	0.7580 (2)	0.7235 (4)	0.7782 (1)
Q33	0.1861 (5)	0.1904 (2)	0.1731 (7)	0.1804 (6)	0.1904 (2)	0.1899 (4)	0.2010 (1)
Q34	0.4156 (6)	0.5451 (1)	0.4633 (3)	0.4156 (6)	0.5451 (1)	0.4224 (4)	0.4203 (5)

Fig. 6 MAP results and their corresponding rank (between brackets) for each query topic in the tolerant scenario. For example, in Q1, the best system was Tika (rank 1), followed by Tika+SymSpell (rank 2). The worst result was obtained by ABBYY + SymSpell (rank 7)

We repeated the same analyses for the strict scenario and the findings remained the same. Those were omitted from the text due to space reasons.

Table 5 computes, for each pair of configurations, how many times the configuration on the column was worse, equivalent, or better than the configuration of the row. For example, we can see that in a comparison with ABBYY, Tornado was worse in 19 topics, equivalent in 10, and better in 5. As done in previous work [5,15], results within a 5% proportional difference were considered equivalent as they can be regarded as unimportant. Looking at the digitization systems, we see again that ABBYY has the most wins and that Tornado was better than Tika (13 wins and 8 losses). SymSpell had proportional improvements of at least 5% in eight topics, compared to Tika and in five compared to ABBYY. sOCRates, on the other hand, had very small changes in comparison with the Tika baseline and no change in relation to ABBYY.

4.4 Intrinsic evaluation

In our intrinsic evaluation, we compare the results of OCR digitization and correction against the ground truth digitizations for a sample of sentences. Table 6 presents the results

Table 5 Pairwise comparisons of all experimental runs

	Tika +sOCRates			Tika +SymSpell			ABBYY			ABBYY +sOCRates			ABBYY +SymSpell			Tornado		
	<	=	>	<	=	>	<	=	>	<	=	>	<	=	>	<	=	>
Tika	1	33	0	2	24	8	2	14	18	0	34	0	3	13	18	8	13	13
+ sOCRates				2	21	11	2	11	21	0	34	0	3	12	19	6	14	14
+ SymSpell							5	14	15	0	34	0	2	17	15	10	12	12
ABBYY										0	34	0	4	25	5	19	10	5
+ sOCRates													4	25	5	19	10	5
+ SymSpell																15	12	7

The cells show the number of topics in which the configuration of the column is worse than (<), equivalent (=), or better than (>) the configuration of the row in terms of MAP. Proportional differences consider a 5% margin

obtained for the error metrics CER and WER and the Pearson correlation with the retrieval quality metrics. Despite the small sample size, the intrinsic analysis confirms the findings of the extrinsic experiments regarding the digitization tools—ABBYY is the best performing, followed by Tornado, and Tika comes last. In a related investigation, [22] compared another set of OCR tools, namely Tesseract, Amazon Textract, and Google Document AI. He found that the best results were provided by the latter. WER scores for the tools ranged between 1.3 and 2.4 for English documents. In the Arabic documents, error rates were much higher, lying between 7.5 and 15.3. Although a direct comparison cannot be made since we are dealing with different documents, our results are closer to the ones obtained for the English documents.

Regarding the correction methods, we see that in many cases they failed to fix problems with the Tika digitizations and inserted more errors. sOCRates results were closer to the ground truth, meaning that fewer errors were inserted. However, this was due to it making fewer changes than SymSpell (as seen in Sect. 4.2, sOCRates does not make changes over ABBYY). This confirms the findings by [55] comparing sOCRates and SymSpell on another dataset. With the Pearson coefficient, we can observe a strong negative correlation between intrinsic error metrics and retrieval quality, meaning that cleaner text yields better retrieval.

As already observed in the extrinsic runs, the SymSpell version used in our experiments has some limitations to handle numbers and special characters. For these cases, an extra step could be added to ignore these types of tokens. Besides this treatment, to improve SymSpell performance, large domain dictionaries with bigrams and unigrams are necessary.

Table 7 presents examples of words extracted by Apache Tika and post-processed by SymSpell and sOCRates. Lines 1–7 show examples of erroneous digitizations (all non-words) in which at least one of the correction systems was able to fix the problems. Lines 8–14 show examples of correct digitizations that had errors inserted by the correction system. In lines 3–5, we see cases in which Tika had problems with hyphenated words that were fixed by SymSpell. On the other hand, sOCRates suggested words that are syntactically similar and more frequent, but incorrect. In line 14, we see an instance of the problem SymSpell had with numbers—numbers were replaced by "de" and "a", which are the most frequent unigrams in our corpus; and ± was dropped.

Comparing the digitizations obtained by the different tools and the ground truth, we were able to identify some common patterns of errors generated by the OCR process. Some common errors were 'I'→'l', 'ô'→'O', 'm'→'rn', and 'fi'→'ft'. These errors can be attributed to their visual similarity, especially depending on the font. Part of the errors are also related to the use of diacritical marks that are common in Portuguese (*e.g.,* ç, á, ã, ó)—we noticed that many mistakes involve such characters. Another issue is to do with REGIS documents spanning a long period of time during which an orthographic reform took place (in 2009). This reform aimed at unifying the spelling across eight Portuguese speaking countries and affected hyphenated words and diacritical marks. These changes pose an additional challenge to OCR and post-OCR tools.

4.5 Comparing with other IR test collections

In order to compare the impacts of OCR correction in another document collection, we reproduced our extrinsic experiments using the CHAVE test collection [47], which is also in Portuguese. While REGIS is composed of large domain-specific documents, CHAVE is composed of short newspaper articles. The average document in REGIS (25K tokens) is about 150 times larger than the average document in CHAVE (with 168 tokens). Since the input documents in CHAVE are already in pure text, we took the version with synthetically created errors by [1] and used by [55].

The same Solr configurations were used in both collections. The baseline run in REGIS is the one in which the digitization system was Apache Tika, and the baseline in CHAVE has synthetic errors inserted in 25% of the words.

Table 6 Intrinsic results and Pearson correlation with retrieval quality metrics (tolerant scenario)

Method	CER	WER	MAP	NDCG	Rel.Ret
Tika	1.22	7.03	0.4947	0.6705	657
Tika + sOCRates	2.55	7.88	0.4904	0.6664	652
Tika + SymSpell	4.47	12.35	0.4810	0.6566	648
ABBYY	**0.40**	**1.62**	0.5438	0.7109	697
ABBYY + sOCRates	0.40	1.64	0.5438	0.7109	697
ABBYY + SymSpell	3.65	9.07	0.5173	0.6869	688
Tornado	2.31	7.60	0.5054	0.6911	666
Pearson correlation		**CER**	−0.7110	−0.7276	−0.5666
		WER	−0.8848	−0.8787	−0.7699

CER and WER are error metrics, so the lower the better

Table 7 Examples of words extracted by the digitization tools (Tika and ABBYY) and their corresponding post-processed versions (by SymSpell and sOCRates)

Tika	+ SymSpell	+ sOCRates
(1) câso ×	caso ✓	câso ×
(2) situaçåo ×	situação ✓	situaçåo ×
(3) conduti- vidade ×	condutividade ✓	conduta cidade ×
(4) consti. tuídos ×	constituídos ✓	consta ruídos ×
(5) Oligoceno-Miocê- nlca ×	Oligoceno Miocênica ✓	Oligocitêmico nuca ×
(6) biocronoestratigráfi co ×	biocronoestratigráfico ✓	biocronoestratigráfi co ×
(7) turbidrticos ×	turbidíticos ✓	turbidíticos ✓
(8) Elmworth ✓	El worth ×	Elmworth ✓
(9) K88–K90 e K100–K130 ✓	Km kM e km de kM de ×	K88–K90 e K100–K130 ✓
(10) SEGEL ✓	hegel ×	SEGEL ✓
(11) (CGMT) mostra ✓	cgt MOStra ×	(CGMT) mostra ✓
(12) sísmica 4D ✓	sísmica D ×	sísmica 4D ✓
(13) bioestratígrafos ✓	bimestre autógrafos ×	bioestratígrafos ✓
(14) 82 ±1 Ma e 48,9 Ma ✓	de a Ma e de a Ma ×	82 ±1 Ma e 48,9 Ma ✓
ABBYY	**+ SymSpell**	**+ sOCRates**
(1) câso ×	caso ✓	câso ×
(2) situaçåo ×	situação ✓	situaçåo ×
(3) conduti-vidade ×	condutividade ✓	conduti-vidade ×
(4) constituídos ✓	constituídos ✓	constituídos ✓
(5) Oligoceno-Miocê-nica ×	Oligoceno Miocênica ✓	Oligoceno-Miocê-nica ×
(6) biocronoestratigráfico ✓	biocronoestratigráfico ✓	biocronoestratigráfico ✓
(7) turbidíticos ✓	turbidíticos ✓	turbidíticos ✓
(8) Elmworth ✓	El worth ×	Elmworth ✓
(9) K88–K90 e K100–K130 ×	Km kM e km de km de ×	K88–K90 e K100–K130 ×
(10) SEGEL ✓	hegel ×	SEGEL ✓
(11) (CGMT) mostra ✓	cGT mostra ×	(CGMT) mostra ✓
(12) sísmica 4D ✓	sísmica D ×	sísmica 4D ✓
(13) bioestratígrafos ✓	bimestre autógrafos ×	bioestratígrafos ✓
(14) 82 ±1 Ma e 48,9 Ma ✓	de a Ma e de a Ma ×	82 ±1 Ma e 48,9 Ma ✓

Correct words have a ✓ and incorrect words have a ×

Table 8 Error correction results obtained in REGIS and CHAVE collections under different configurations

Method	MAP	Δ	PR@10	Δ	Rel.Ret	Δ	NDCG	Δ
REGIS Tika								
Baseline	.4947	–	.6294	–	657	–	.6705	–
+ SymSpell	.4810	−2.77%	.6265	−0.46%	648	−1.37%	.6566	−2.07%
+ sOCRates	.4904	−0.87%	.6147	−2.34%	652	−0.76%	.6664	−0.61%
CHAVE (Synthetic)								
Baseline	.2156	–	.2330	–	749	–	.3653	–
+ SymSpell	.2952	36.92%	.3280	40.77%	1106	47.66%	.4778	30.80%
+ sOCRates	.2657	23.24%	.3091	32.66%	1079	44.06%	.4459	22.06%
Ideal	.3075	42.63%	.3290	41.20%	1139	52.07%	.4891	33.89%

In CHAVE, we also have an *ideal* run with the original clean texts that works as an upper bound for the correction systems. Table 8 presents the results of this comparative experiment. We can see that, unlike REGIS, on CHAVE both correction systems improved results in all metrics. We attribute this difference to three reasons (*i*) the size of the documents: it seems well established in the literature that shorter documents may benefit more from correction [10,51]; (*ii*) the error rates in the baseline run in CHAVE were significantly higher—*i.e.,* a 25% WER compared to a 7.03% WER in REGIS, which mean a larger room for improvements; and (*iii*) CHAVE consists of news articles with a simpler language that is mostly covered within the lexica used in the post-OCR systems. Thus, it is easier to correct compared to a technical collection such as REGIS.

5 Conclusion

In this article, we analyzed the impacts of OCR digitization and error correction in information retrieval. Our experiments were done in a collection of PDF documents in Portuguese. The textual contents of the documents were extracted using three tools (ABBYY, Tika, and Tornado) and post-processed by two correction methods (sOCRates and SymSpell). Next, we discuss the implication of our work regarding our research questions.

Our experiments with three OCR tools (Sect. 4.1) found statistically significant differences in retrieval accuracy metrics. ABBYY, the best performer, yielded MAP scores that were about 4 percentage points higher than Tika's. The documents in our experiments are typically very long (thesis, dissertations, and technical reports, averaging 25K tokens), and yet they were significantly impacted by OCR errors. Unlike existing work that suggested that long documents were robust to these errors [10,39,51], here we showed that even very long documents are impacted by OCR errors.

Existing work on the impact of OCR error correction on IR metrics has reached contrasting conclusions. While some works pointed out that post-processing can improve IR quality [14,55,57] others found the opposite [10,39,51]. In our experiments, we found that, on the average for the complete set of query topics, error correction did not help. All retrieval metrics were lower than for the baseline run—but the differences were not statistically significant. Then, in a topic-by-topic analysis SymSpell was able to improve retrieval results in 19 out of 34 topics. These results are also in opposition to previous work that suggested that long documents would not benefit from OCR correction. Only two query topics had severe performance drops with the correction by SymSpell, and both were due to the same specific issue of how SymSpell deals with numbers. Nevertheless, the quality of the correction systems needs improvements before they can be used to automatically process the outputs of OCR.

We also carried out an intrinsic evaluation to calculate OCR error rate metrics (Sect. 4.4). That required the manual creation of a ground truth for the digitization process. The results also confirm that ABBYY was the best digitization tool with an estimated word error rate of 1.62, while Tika and Tornado obtained an error rate four times greater. Still, budget constraints may deter the adoption of paid tools such as ABBYY, Amazon Textract, or Google Document AI. [3] identified several factors that affect digitization costs, which include the expected quality of the output. According to the figures reported by [22], it would cost between 3600 and 145,000 US dollars to process the entire REGIS collection with these cloud tools.[21] These costs, allied to the fact that Tika is free and can be easily integrated into one's code, mean it will be the tool of choice in many practical applications, despite its higher error rates. Consequently, error correction methods are likely to continue to be needed.

A limitation in our intrinsic evaluation is the small number of samples in our ground truth dataset. Although these experiments allowed important insights, the ground truth is not a representative sample for REGIS collection. Extra annota-

[21] REGIS documents have a total of 2.4 million pages. The costs mentioned by [22] range between $1.5 and 60 US dollars per 1000 pages.

tion effort would be necessary to complement this sample and allow a more accurate evaluation. Additionally, the documents in REGIS contain not only text, but also figures, tables, and equations. We let the evaluation of the impact of dealing with these elements for future work.

In this article, we used only one scoring function (BM25) and it is possible that other retrieval techniques could be more (or less) impacted by OCR error and error correction. Along the same lines, new IR techniques were proposed in recent years—neural document reranking [43], in special, has shown significant improvements in retrieval quality. It would be interesting to investigate whether these techniques are able to make IR systems more robust to OCR errors.

Acknowledgements The authors thank the anonymous reviewers whose suggestions helped us improve our manuscript. We also thank Moniele K. Santos for her help in creating the ground truth. This work was partially supported by Petrobras 2017/00752-3, CAPES Finance Code 001, and CNPq/Brazil. The authors acknowledge the National Laboratory for Scientific Computing (LNCC/MCTI, Brazil) for providing HPC resources of the SDumont supercomputer, which have contributed to the research results reported within this article (URL: http://sdumont.lncc.br).

Author Contributions LLdO was involved in the conceptualization, methodology, software, investigation, writing—original draft, and visualization. DSV contributed to the methodology, software, and data curation. AMAA helped in the conceptualization, software, and writing–review and editing. FCC assisted in the conceptualization, supervision, writing–review and editing, and funding acquisition. DdSMG contributed to the conceptualization and writing–review and editing. MCR assisted in the conceptualization and writing—review and editing. RKR performed the conceptualization and writing—review and editing. VPM contributed to the conceptualization, methodology, writing—original draft, writing–review and editing and project administration.

Code Availability The code we implemented to run the experiments in this article is available at https://github.com/lucaslioli/solr-query-script. The datasets generated during and analyzed during the current study are available in https://github.com/Petroles/regis-collection and https://github.com/lucaslioli/regis-collection-gs.

Declarations

Conflict of Interest The authors have no competing interests to declare.

References

1. Bazzo, G.T., Lorentz, G.A., Vargas, D.S., et al.: Assessing the impact of OCR errors in information retrieval. In: European Conference on Information Retrieval, pp. 102–109 (2020)
2. Bender, E.M.: On achieving and evaluating language-independence in nlp. Linguist. Issues Lang. Technol. **6** (2011)
3. Bia, A., Muñoz, R., Gómez, J.: DiCoMo: the digitization cost model. Int. J. Digital Lib. **11**(2), 141–153 (2010)
4. Boros, E., Nguyen, N.K., Lejeune, G., et al.: Assessing the impact of OCR noise on multilingual event detection over digitised documents. Int. J. Digital Lib. pp. 1–26 (2022)
5. Buckley, C., Voorhees, E.M.: Evaluating evaluation measure stability. In: ACM SIGIR Forum, pp. 235–242 (2017)
6. Carrasco, R.C.: An open-source OCR evaluation tool. In: Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, pp. 179–184 (2014)
7. Castro, J.D.B., Canchumuni, S.W.A., Villalobos, C.E.M., et al.: Improvement optical character recognition for structured documents using generative adversarial networks. In: 2021 21st International Conference on Computational Science and Its Applications (ICCSA), pp. 285–292 (2021)
8. Chiron, G., Doucet, A., Coustaty, M., et al: ICDAR2017 competition on post-OCR text correction. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp .1423–1428 (2017)
9. Consoli, B., Santos, J., Gomes, D., et al.: Embeddings for named entity recognition in geoscience portuguese literature. In: Proceedings of The 12th Language Resources and Evaluation Conference, pp. 4625–4630 (2020)
10. Croft, W.B., Harding, S., Taghva, K., et al.: An evaluation of information retrieval accuracy with simulated OCR output. In: Symposium on Document Analysis and Information Retrieval, pp. 115–126 (1994)
11. Drobac, S., Lindén, K.: Optical character recognition with neural networks and post-correction with finite state methods. Int. J. Document Anal. Recog. (IJDAR) **23**(4), 279–295 (2020)
12. Dutta, H., Gupta, A.: PNRank: Unsupervised ranking of person name entities from noisy OCR text. Decis. Support Syst. **152**(113), 662 (2022)
13. Ehrmann, M., Hamdi, A., Pontes, E.L., et al.: Named entity recognition and classification on historical documents: A survey. arXiv preprint arXiv:2109.11406 (2021)
14. Evershed, J., Fitch, K.: Correcting noisy OCR: Context beats confusion. In: Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, pp. 45–51 (2014)
15. Flores, F.N., Moreira, V.P.: Assessing the impact of stemming accuracy on information retrieval-a multilingual perspective. Inf. Process. Manag. **52**(5), 840–854 (2016)
16. Francois, M., Eglin, V., Biou, M.: Text detection and post-OCR correction in engineering documents. In: Uchida, S., Barney, E., Eglin, V. (eds.) Document Analysis Systems, pp. 726–740. Springer International Publishing, Cham (2022)
17. Ghosh, K., Chakraborty, A., Parui, S.K., et al.: Improving information retrieval performance on OCRed text in the absence of clean text ground truth. Inf. Process. Manag. **52**(5), 873–884 (2016)
18. Gomes, D., Cordeiro, F., Consoli, B., et al.: Portuguese word embeddings for the oil and gas industry: Development and evaluation. Comput. Ind. **124**(103), 347 (2021)
19. Gupte, A., Romanov, A., Mantravadi, S., et al.: Lights, camera, action! a framework to improve nlp accuracy over OCR documents (2021)
20. Hämäläinen, M., Hengchen, S.: From the Paft to the Fiiture: a Fully Automatic NMT and Word Embeddings Method for OCR Post-Correction. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pp. 431–436 (2019)
21. Hamdi, A., Jean-Caurant, A., Sidère, N., et al.: Assessing and minimizing the impact of OCR quality on named entity recognition. In: International Conference on Theory and Practice of Digital Libraries, Springer, pp. 87–101 (2020)
22. Hegghammer, T.: OCR with tesseract, amazon textract, and google document ai: a benchmarking experiment. J. Comput. Social Sci. 1–22 (2021)
23. Hull, D.: Using statistical testing in the evaluation of retrieval experiments. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 329–338 (1993)

24. Huynh, V.N., Hamdi, A., Doucet, A.: When to use OCR post-correction for named entity recognition? In: International Conference on Asian Digital Libraries, Springer, pp. 33–42 (2020)
25. Jiang, M., Hu, Y., Worthey, G., et al.: Impact of OCR quality on BERT embeddings in the domain classification of book excerpts. Proceedings http://ceur-ws.org ISSN 1613:0073 (2021)
26. Jing, H., Lopresti, D., Shih, C.: Summarization of noisy documents: A pilot study. In: Proceedings of the HLT-NAACL 03 text summarization workshop, pp. 25–32 (2003)
27. Johnson, S., Jourlin, P., Jones, K.S., et al.: Spoken document retrieval for TREC-7 at cambridge university. In: TREC, p. 1 (1999)
28. Kantor, P.B., Voorhees, E.M.: The TREC-5 confusion track: Comparing retrieval methods for scanned text. Inf. Retrieval **2**(2), 165–176 (2000)
29. Kettunen, K., Keskustalo, H., Kumpulainen, S., et al.: OCR quality affects perceived usefulness of historical newspaper clippings–a user study (2022). https://arxiv.org/abs/2203.03557
30. Lam-Adesina, A.M., Jones, G.J.: Examining and improving the effectiveness of relevance feedback for retrieval of scanned text documents. Inf. Process. Manag. **42**(3), 633–649 (2006)
31. Lawley, C.J., Raimondo, S., Chen, T., et al.: Geoscience language models and their intrinsic evaluation. Appl. Comput. Geosci., 100084 (2022)
32. Lin, X.: Impact of imperfect OCR on part-of-speech tagging. In: Seventh International Conference on Document Analysis and Recognition, Proceedings., pp. 284–288 (2003)
33. Linhares Pontes, E., Hamdi, A., Sidere, N., et al.: Impact of OCR quality on named entity linking. In: International Conference on Asian Digital Libraries, Springer, pp. 102–115 (2019)
34. Linhares Pontes, E., Cabrera-Diego, L.A., Moreno, J.G., et al.: MELHISSA: a multilingual entity linking architecture for historical press articles. Int. J. Digital Lib. 1–28 (2021)
35. Ma, X., Pradeep, R., Nogueira, R., et al.: Document expansion baselines and learned sparse lexical representations for ms marco v1 and v2. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 3187–3197 (2022)
36. Martínek, J., Lenc, L., Král, P.: Building an efficient OCR system for historical documents with little training data. Neural Comput. Appl. **32**(23), 17,209-17,227 (2020)
37. Mei, J., Islam, A., Moh'd, A., et al.: Statistical learning for OCR error correction. Inf. Process. Manag. **54**(6), 874–887 (2018)
38. Miller, D., Boisen, S., Schwartz, R., et al.: Named entity extraction from noisy input: speech and OCR. In: Sixth Applied Natural Language Processing Conference, pp. 316–324 (2000)
39. Mittendorf, E., Schäuble, P.: Information retrieval can cope with many errors. Inf. Retrieval **3**(3), 189–216 (2000)
40. Mutuvi, S., Doucet, A., Odeo, M., et al.: Evaluating the impact of OCR errors on topic modeling. In: International Conference on Asian Digital Libraries, pp. 3–14 (2018)
41. Nguyen, T., Jatowt, A., Coustaty, M., et al.: Deep statistical analysis of OCR errors for effective post-OCR processing. In: Joint Conference on Digital Libraries (JCDL), pp. 29–38 (2019)
42. Nguyen, T.T.H., Jatowt, A., Coustaty, M., et al.: Survey of post-OCR processing approaches. ACM Comput. Surv. (CSUR) **54**(6), 1–37 (2021)
43. Nogueira, R., Cho, K.: Passage re-ranking with bert. arXiv preprint arXiv:1901.04085 (2019)
44. Lima de Oliveira, L., Romeu, R.K., Moreira, V.P.: REGIS: A test collection for geoscientific documents in portuguese. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2363–2368 (2021)
45. Rigaud, C., Doucet, A., Coustaty, M., et al.: ICDAR 2019 competition on post-OCR text correction. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1588–1593 (2019)
46. Sakai, T.: Statistical reform in information retrieval? In: ACM SIGIR Forum, pp. 3–12 (2014)
47. Santos, D., Rocha, P.: The key to the first CLEF with portuguese: Topics, questions and answers in CHAVE. In: Workshop of the Cross-Language Evaluation Forum for European Languages, pp. 821–832 (2004)
48. Singh, S.: Optical character recognition techniques: a survey. J. Emerg. Trends Comput. Inf. Sci. **4**(6), 545–550 (2013)
49. Smucker, M.D., Allan, J., Carterette, B.: A comparison of statistical significance tests for information retrieval evaluation. In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, pp. 623–632 (2007)
50. van Strien, D., Beelen, K., Ardanuy, M.C., et al.: Assessing the impact of OCR quality on downstream NLP tasks. In: Proceedings of the 12th International Conference on Agents and Artificial Intelligence, ICAART, pp. 484–496 (2020)
51. Taghva, K., Borsack, J., Condit, A., et al.: The effects of noisy data on text retrieval. J. Am. Soc. Inf. Sci. **45**(1), 50–58 (1994)
52. Taghva, K., Borsack, J., Condit, A.: Effects of OCR errors on ranking and feedback using the vector space model. Inf. Process. Manag. **32**(3), 317–327 (1996)
53. Taghva, K., Borsack, J., Condit, A.: Evaluation of model-based retrieval effectiveness with OCR text. ACM Trans. Inf. Syst. (TOIS) **14**(1), 64–93 (1996)
54. Traub, M.C., Samar, T., Van Ossenbruggen, J., et al.: Impact of crowdsourcing OCR improvements on retrievability bias. In: Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, pp. 29–36 (2018)
55. Vargas, D.S., de Oliveira, L.L., Moreira, V.P., et al.: sOCRates-a post-OCR text correction method. In: Anais do XXXVI Simpósio Brasileiro de Bancos de Dados, pp. 61–72 (2021)
56. Wiedenhofer, L., Hein, H.G., Dengel, A.: Post-processing of OCR results for automatic indexing. In: Proceedings of 3rd International Conference on Document Analysis and Recognition, IEEE, pp. 592–596 (1995)
57. Zhuang, S., Zuccon, G.: Dealing with typos for BERT-based passage retrieval and ranking. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 2836–2842 (2021)
58. Zosa, E., Mutuvi, S., Granroth-Wilding, M., et al.: Evaluating the robustness of embedding-based topic models to ocr noise. In: International Conference on Asian Digital Libraries, Springer, pp. 392–400 (2021)
59. Zu, G., Murata, M., Ohyama, W., et al.: The impact of OCR accuracy on automatic text classification. In: Advanced Workshop on Content Computing, pp. 403–409 (2004)