UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

LUCAS LIMA DE OLIVEIRA

Creating Resources and Evaluating the Impact of OCR Quality on Information Retrieval: A Case Study in the Geoscientific Domain

Thesis presented in partial fulfillment of the requirements for the degree of Master of Computer Science

Advisor: Prof[a]. Dr[a]. Viviane Pereira Moreira

Porto Alegre
March 2022

CIP — CATALOGING-IN-PUBLICATION

Oliveira, Lucas Lima de

Creating Resources and Evaluating the Impact of OCR Quality on Information Retrieval: A Case Study in the Geoscientific Domain / Lucas Lima de Oliveira. – Porto Alegre: PPGC da UFRGS, 2022.

64 f.: il.

Thesis (Master) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2022. Advisor: Viviane Pereira Moreira.

1. Information retrieval. 2. Test collection. 3. OCR errors. 4. Error correction. I. Moreira, Viviane Pereira. II. Título.

“A ship in port is safe,
but that’s not what ships are built for.
Sail out to sea and do new things”
— Grace Hopper

ACKNOWLEDGMENTS

I want to firstly thank my advisor, Prof[a]. Dr[a]. Viviane Moreira, who supported me, collaborated, and entirely participated in my research. Thank you for your advice, patience, friendship, and mainly, thank you so much for believing and giving me this opportunity. During these years, I learned a lot and evolved personally and professionally.

This work also would not have been possible without the help of my colleagues. Thank you all for the knowledge exchange, collaboration, feedback, and for helping me to accomplish my research.

I also would like to thank my family and friends, especially my mother, Siminea Lima, and my partner, Amanda Rodrigues, who were by my side all the time. Their support was essential, especially during working days until late at night. Special mentions go to my close friends Gregory Fontoura, Iago Corrêa, and Mateus Berndt, who are far from being merely friends and are like brothers.

Additionally, I would like to thank e Federal University of Rio Grande do Sul and the Institute of Informatics for providing everything that was needed to complete my research. This work was partially supported by Petrobras 2017/00752-3, CAPES Finance Code 001, and CNPq/Brazil. I 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 work (URL: <http://sdumont.lncc.br>).

ABSTRACT

The Portable Document Format (PDF) has become the de facto standard for document storage and sharing. Scientific papers, project proposals, contracts, books, legal documents are typically stored and distributed as PDF files. While extracting the textual contents of born-digital PDF documents can be done with high accuracy, if the document consists of a scanned image, Optical Character Recognition (OCR) is typically required. The output of OCR can be noisy, especially when the quality of the scanned image is poor – really common on historical documents –, which in turn can impact downstream tasks such as Information Retrieval (IR). Post-processing OCR-ed documents is an alternative to fix extraction errors and, intuitively, improve the results of downstream tasks. This work evaluates the impact of OCR extraction and correction on IR. We compared different extraction and correction methods on OCR-ed data from real scanned documents. To evaluate IR tasks, the standard paradigm requires a test collection with documents, queries, and relevance judgments. Creating test collections requires significant human effort, mainly for providing relevance judgments. As a result, there are still many domains and languages that, to this day, lack a proper evaluation testbed. Portuguese is an example of a major world language that has been overlooked in terms of IR research – the only test collection available is composed of news articles from 1994 and a hundred queries. With the aim of bridging this gap, we developed REGIS (Retrieval Evaluation for Geoscientific Information Systems), a test collection for the geoscientific domain in Portuguese. REGIS contains 20K documents and 34 query topics along with relevance assessments. Our results from the experiments with REGIS showed that on average for the complete set of query topics, retrieval quality metrics change very little. However, a more detailed analysis revealed that most query topics improved with error correction.

Keywords: Information retrieval. test collection. OCR errors. error correction.

Criando Recursos e Avaliando o Impacto da Qualidade do OCR na Recuperação da Informação: Um Estudo de Caso no Domínio Geocientífico

RESUMO

O Formato de Documento Portátil (PDF) se tornou um dos padrões mais usados para armazenamento e compartilhamento de documentos. Artigos científicos, propostas de projetos, contratos, livros e documentos jurídicos são normalmente armazenados e distribuídos como arquivos PDF. Embora a extração do conteúdo textual de documentos PDF originados de forma digital possa ser feita com alta precisão, se o documento consistir em uma imagem digitalizada, o Reconhecimento Óptico de Caracteres (OCR) é normalmente necessário. A saída do OCR pode ser ruidosa, especialmente quando a qualidade da imagem digitalizada é ruim – muito comum em documentos históricos –, o que por sua vez pode impactar tarefas posteriores, como Recuperação de Informação (IR). O pós-processamento de documentos OCR é uma alternativa para corrigir erros de extração e, intuitivamente, melhorar os resultados em tarefas posteriores. Este trabalho avalia o impacto da extração e correção de OCR em IR. Comparamos diferentes métodos de extração e correção em textos extraídos por OCR de documentos escaneados reais. Para avaliar as tarefas de IR, o paradigma padrão requer uma coleção de testes com documentos, consultas e julgamentos de relevância. A criação de coleções de teste requer um esforço humano significativo, principalmente na realização dos julgamentos de relevância. Como resultado, ainda existem muitos domínios e idiomas que, até hoje, carecem de um ambiente de teste para avaliação adequada. O português é um exemplo de uma importante língua mundial que tem sido negligenciada em termos de pesquisas de IR - a única coleção de testes disponível é composta por notícias de 1994 e uma centena de consultas. Com o objetivo de preencher essa lacuna, desenvolvemos a REGIS (*Retrieval Evaluation for Geoscientific Information Systems*), uma coleção de testes para o domínio geocientífico em português. REGIS contém 20 mil documentos e 34 tópicos de consulta, juntamente com julgamentos de relevância. Nossos resultados dos experimentos utilizando a REGIS mostraram que, em média, para o conjunto completo de tópicos de consulta, as métricas de qualidade de recuperação variam muito pouco. No entanto, uma análise mais detalhada revelou que a maioria dos tópicos de consulta melhorou com a correção de erros.

Palavras-chave: Recuperação de informações. colecões de teste. Avaliação de OCR.

LIST OF ABBREVIATIONS AND ACRONYMS

ANN	Artificial Neural Networks
CER	Character Error Rate
BM25	Best Matching 25
DFR	Divergence From Randomness
IBICT	Brazilian Institute of Information in Science and Technology
IR	Information Retrieval
ICDAR	International Conference on Document Analysis and Recognition
MAP	Mean Average Precision
NDCG	Normalized Discounted Cumulative Gain
NLP	Natural Language Processing
OCR	Optical Character Recognition
PDF	Portable Document Format
QLD	Query Likelihood with Dirichlet smoothing
REGIS	Retrieval Evaluation for Geoscientific Information Systems
RM3	Relevance Model 3
XML	Extensible Markup Language
WER	Word Error Rate

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

A significant part of textual information exchanged in digital documents, such as scientific articles, thesis, technical reports, project proposals and contracts, is typically stored and distributed in Portable Document Format (PDF). A report from the PDF association[1] 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 files sent as e-mail attachments in Outlook Exchange Enterprise were in PDF.

The digitization of industries highly increased the usage of PDF files and brought many new challenges to information retrieval (IR) topic, including dealing with large volumes of data, multiple modalities, and multiple languages. These problems have been the focus of much research over the years and solutions to them were proposed. To test the efficiency of any IR solution, a test collection is fundamental.

Given their importance, significant effort was devoted to building test collections since the early days of IR research (CLEVERDON, 1962). These efforts were intensified in the 90s in the US, with the Text REtrieval Conferences (TREC)[2]. Similar efforts in Europe (CLEF)[3] and Asia (NTCIR)[4] also emerged. Several test collections were created within the scope of these evaluation campaigns, addressing different retrieval tasks and languages. Yet, the cost of creating this type of resource means there are still many domains and languages that, to this day, lack a proper evaluation testbed.

Portuguese is an example of a major world language (the sixth-largest language with over 228 million native speakers across four continents) that has been overlooked in terms of linguistic resources. The only existing IR test collection was created in the CLEF evaluation campaigns and consists of news documents published by Folha de São Paulo and Público (newspapers from Brazil and Portugal, respectively) from 1994 and 1995. The collection has 100 queries with relevance judgments (SANTOS; ROCHA, 2004).

The Oil and Gas (O&G) industry plays an important role in Portuguese-speaking countries, representing an essential part of their economies. With the discovery of the Brazilian pre-salt and recent investments on it, many exploration and production projects have emerged. As pointed out by Gomes et al. (2021), despite the importance of this in-

[1]PDF Association: <https://www.pdfa.org/wp-content/uploads/2018/06/1330_Johnson.pdf>
[2]TREC: <https://trec.nist.gov/>
[3]CLEF: <http://www.clef-initiative.eu/>
[4]NTCIR: <http://ntcir.nii.ac.jp/>

dustry, there are few linguistic resources available for this sub-domain of the Geosciences. O&G companies deal with many types of unstructured textual information, including technical, geoscientific, and production reports, scientific papers, thesis, operational logs, and analyses (GOMES et al., 2021). Looking at the case of the biggest Brazilian oil company, Petrobras, 94% of the textual data is represented in PDF files. These documents are characterized by having many elements such as maps, graphs, figures, tables, and formulas. In addition, they are commonly very long and collected over a long period of time, which can also include spelling variations. These documents also incorporate a highly technical vocabulary (*e.g.*, names of basins, fields, rocks, geological ages, *etc.*), which is not covered by the available IR test collections, especially in Portuguese. This limitations poses difficulties for Brazilian researchers to test their methods and for companies to implement them internally.

In an attempt to address this gap, we created a test collection for the geoscientific domain in Portuguese. The collection is called REGIS[5] (Retrieval Evaluation for Geoscientific Information Systems); it is composed of over 20 thousand documents, 34 query topics, and their corresponding relevance judgments. The documents were produced over a long time span (1957 to 2020) and vary substantially in terms of visual quality. REGIS was created with the cooperation of domain specialists, following the pooling method proposed by Spark-Jones (1975) and well described by Sanderson (2010).

Before being fed to Natural Language Processing (NLP) or 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 extraction process involves the use of Optical Character Recognition (OCR) algorithms to identify the textual elements within the image.

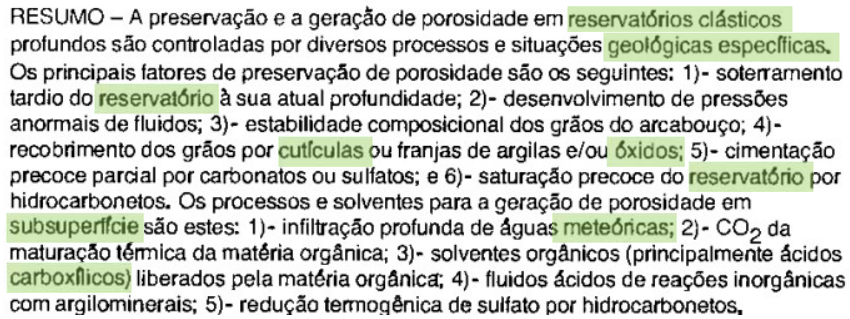
Although OCR technology has been improving over the years, it is still not perfect. Furthermore, the quality of scanned text may be poor, especially for older documents. Singh (2013) estimated that, with an accuracy rate of 99% at the character level, assuming an average word length of five characters, means that one in 20 words would have an extraction error (*i.e.,* a 5% word error rate). Bazzo et al. (2020) demonstrated that starting at a 5% word error rate, significant impacts are noticed in retrieval quality.

Figure 1.1 shows an excerpt of a real OCR extraction error in a document from the REGIS collection. The original PDF document, Figure 1.1 (a), was processed by Apache Tika to extract its textual contents, which are shown in Figure 1.1 (b). Extraction errors are highlighted in yellow and their counterparts in the original PDF are in green. All

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

occurrences of *reservatório* (reservoir) were erroneously extracted to *reservat6rio*. As a result, queries with the keyword *reservatório* would not be able to retrieve the 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)[6], in line 1, incorrectly extracted as *elásticos* (elastic), which is a valid word in Portuguese with a completely different meaning. Bazzo et al. (2020) 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 works with new approaches for post-OCR text correction (MEI et al., 2018; HÄMÄLÄINEN; HENGCHEN, 2019; DROBAC; LINDÉN, 2020; VARGAS et al., 2021).

Figure 1.1 – Example of OCR errors in a document from the REGIS collection


RESUMO – A preservação e a geração de porosidade em reservatórios clásticos profundos são controladas por diversos processos e situações geológicas especfficas. Os principais fatores de preservação de porosidade são os seguintes: 1)- soterramento tardio do reservatório à sua atual profundidade; 2)- desenvolvimento de pressões anormais de fluidos; 3)- estabilidade composicional dos grãos do arcabouço; 4)- recobrimento dos grãos por cutículas ou franjas de argilas e/ou óxidos; 5)- cimentação precoce parcial por carbonatos ou sulfatos; e 6)- saturação precoce do reservatório por hidrocarbonetos. Os processos e solventes para a geração de porosidade em subsuperffcie são estes: 1)- infiltração profunda de águas meteóricas; 2)- CO_2 da maturação térmica da matéria orgânica; 3)- solventes orgânicos (principalmente ácidos carboxílicos) liberados pela matéria orgânica; 4)- fluidos ácidos de reações inorgânicas com argilominerais; 5)- redução termogênica de sulfato por hidrocarbonetos.

(a) Original PDF Document

(b) Extracted Textual Contents

Source: The Authors

Although the impacts of OCR-ed text in IR have already been studied (BAZZO et al., 2020; GHOSH et al., 2016; CROFT et al., 1994; TAGHVA; BORSACK; CONDIT, 1996b; KANTOR; VOORHEES, 2000), there is not much work on evaluating the impact of post-processing techniques that try to fix extraction errors. Our earlier work (VARGAS

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

et al., 2021) 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. However, extraction errors present distinct patterns across different languages, demanding language-specific solutions. Therefore, it is crucial to conduct further research to build resources and develop solutions that can address languages other than English (BENDER, 2019). In a recent survey on this topic, Nguyen et al. (2021) concludes by stating 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 (RIGAUD et al., 2019), datasets in 10 European languages (Bulgarian, Czech, Dutch, English, Finish, French, German, Polish, Spanish, and Slovak) were made available. With the creation of REGIS, in this work, our target language is Portuguese, which despite being the 6^{th} language in the number of native speakers, does not count with a public dataset for evaluating real OCR errors.

With the REGIS collection, we performed experiments to evaluate three text extraction tools with OCR capabilities and two post-processing methods. More specifically, we aim to answer two main research questions: (i) How does the quality of the text extraction affect retrieval results? and (ii) Can post-processing OCR-ed texts improve retrieval quality? Contrary to existing work that argue that long documents are robust to OCR errors, we found that retrieval quality metrics varied significantly depending on the extraction 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 work are:

- Creation of an annotation system to obtain relevant judgments.
- Creation of a new Portuguese IR test collection.
- An investigation of the impact of different text extraction and correction methods for OCR-ed texts using real OCR-ed data.
- An evaluation of the intrinsic quality of text extraction and error correction.
- Experiments with a language that, despite being widely spoken, is underrepresented in terms of IR resources.

The remainder of this document is organized as follows. Section 2 introduces the background with fundamental concepts of text extraction and test collections. Chapter 3 discuss related works. Chapter 4 presents the test collection REGIS, detailing the entire creation process. Chapter 5 describes the different materials and methods used in our experiments. Section 6 reports the results obtained with all experiments. Finally, Chapter 7 concludes this work and discusses future directions.

2 BACKGROUND

This chapter introduces the fundamental concepts about the two main topics covered in this work, namely the creation of test collections and the process of OCR extraction. These are addressed in the next sections.

2.1 Information Retrieval

Information Retrieval (IR) is widely present in the daily lives of people that have access to the Internet. Modern Web search engines such as Google and Bing make use of IR mechanisms to retrieve web pages in response to user queries. Manning, Raghavan and Schütze (2008) defines IR as the process of finding relevant documents in a large collection of unstructured data (usually texts) that satisfy an information need. In this context, the information need is a topic of interest that the user wishes to obtain relevant materials about (i.e., to retrieve these documents). The users express their needs in the format of a short query composed of a few keywords.

Evaluating retrieval quality is crucial for IR and has been a topic of interest since the 1960's. Evaluation requires test collections and quality metrics which are described in the next sections.

2.1.1 Test Collections

Test collections are fundamental resources to evaluate the quality and effectiveness of IR systems. To measure the performance in retrieval tasks, test collections should have three main components: (i) a set of documents, (ii) a set of query topics, and (iii) a set of relevance judgments for the query-document pairs (MANNING; RAGHAVAN; SCHüTZE, 2008; SANDERSON, 2010). The most accepted and employed methodology for creating test collections is the *pooling method* proposed by Spark-Jones (1975). More recently, Sanderson (2010) presented a series of recommendations about the creation of these collections and the evaluation of IR systems.

Since the generation of relevance judgments for query-document pairs is unfeasible, the pooling method generates a subset (pool) of possibly relevant documents for each query. To avoid system bias, Sanderson (2010) recommends that more than one score

function should be used to generate a rank with n (usually $n = 100$ or 50) documents. Then, the best-ranked ones in all systems will be part of the pool, as the most likely documents to be relevant to the query. This method reduces the manual effort needed to annotate all documents in the collection, to only 100 by query.

This methodology has been responsible for the construction of many large test collections and its robustness has already been proven. Even with incomplete relevance judgments, where some possible relevant documents are left out of the pool, previous works (BUCKLEY; VOORHEES, 2017; BUCKLEY; VOORHEES, 2004; YILMAZ; ASLAM, 2006), have shown that some retrieval quality metrics are robust and stable to handle this scenario.

Observing different test collections the structure of the documents and query topics may vary. For the documents, beyond the contents, the only required field is a unique identifier. And in the case of the query topics, the most commonly find components are:

- Topic unique identifier.
- A short title, which is commonly used as the default query submitted to the IR system.
- A short description of the information need, generally, with no more than one sentence.
- A more detailed narrative that helps the annotator decide on the relevance of the documents.

2.1.2 Retrieval Quality Metrics

To evaluate IR systems, the main metrics adopted are based on precision and recall. Precision (Equation 2.1) is the fraction of the retrieved documents that are indeed relevant. Recall (Equation 2.2) is the fraction of the relevant documents that were retrieved. Those metrics are intended to evaluate sets, so they are not suitable for evaluating rankings. As a result, a set of metrics that take into consideration the position in which the documents appear in the ranking are typically used.

Precision at 10 (PR@10) is also widely used, measuring the precision among the top ten documents in the ranking, and the precision is calculated as in Equation. 2.1.

$$Precision = \frac{\#(relevant\ items\ retrieved)}{\#(retrieved\ items)} \tag{2.1}$$

$$Recall = \frac{\#(relevant\ items\ retrieved)}{\#(relevant\ items)} \tag{2.2}$$

The Mean average precision (MAP) is the standard evaluation metric when the relevance judgments are binary as it has a good discrimination power and stability (MANNING; RAGHAVAN; SCHüTZE, 2008). MAP (Equation. 2.3), as the name describes, is calculated as the general mean from average precision of all queries, providing a single measure of quality across recall levels (MANNING; RAGHAVAN; SCHüTZE, 2008).

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{d_j} \sum_{k=1}^{d_j} Precision(R_{jk}) \tag{2.3}$$

where Q is the set of queries, j is the query, R_{jk} corresponds to the set of ranked retrieval results from the top result until the document d_k. When a relevant document is not retrieved, $Precision(R_{jk})$ is taken to be 0.

When the relevance judgments are non-binary, then the Normalized Discounted Cumulative Gain (*NDCG*) can be used. This metric takes into account the relevance levels of the documents to a specific query. NDCG value is calculated as in Equation 2.4

$$NDCG(Q, k) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} Z_{kj} \sum_{d=1}^{k} \frac{2^{R(j,d)} - 1}{log_2(1 + d)} \tag{2.4}$$

where Q is the set of queries, $R(j, d)$ corresponds to the relevance score assessors gave to document d for query j. Z_{kj} is a normalization factor so that a perfect ranking's NDCG at k for query j is 1.

2.2 OCR

OCR is the process of automatically translating, or extracting, text present in digital images (SINGH, 2013). In the IR field, this process is commonly used to process image or PDF files and generate text documents to be indexed. 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, which can be a problem for IR systems. Nowadays, it has still been the topic for researches, such as the very recent work by Nguyen et al. (2021), which surveys a study about OCR quality and post-OCR processing approaches.

2.2.1 Text Extraction Methods

OCR methods face many challenges, such as stylized fonts, figures, scanned pages with noisy signals, multiple languages, document formats, *etc*. Most OCR systems perform image treatment, applying filters aiming to reduce the noise and easily recognize features, such as the borders and gaps between words. Memon et al. (2020) published a recent review on handwritten OCR and showed that the number of publications in this topic increased in the latest years. The extraction methods are divided into five categories: Artificial Neural Networks (ANN), kernel-based, statistical, template matching, and pattern recognition.

According to Memon et al. (2020), from these five main approaches, with the popularization of deep neural architectures, ANN methods, such as Recurrent Neural Network (RNN) and Convolutional Neural Networks (CNN), have become the best techniques for recognition tasks. Older ANN methods were based on feedforward networks, mainly Multi-Layer Perceptron (MLP). Other powerful approaches are kernel-based. In this group, there are models such as Support Vector Machine (SVM), Kernel Fisher Discriminant Analysis (KFDA), and Kernel Principal Component Analysis (KPCA).

Statistical methods are divided into parametric and non-parametric. Examples of the first group are Logistic Regression (LR), Linear Discriminant Analysis (LDA), and Hidden Markov Model (HMM). For the second group, we have K Nearest Neighbors (KNN) and Decision Trees (DT). The advantage of parametric classifiers is that they are faster to learn and can be trained with a small set, while non-parametric are more flexible in learning.

The fourth technique described by Memon et al. (2020) is template matching, which consists in comparing image pieces and predefined templates, as the name suggests. The matching is based on similarity functions, *e.g.,* Euclidean distance, cross-correlation, and normalized correlation. Finally, the last group, structural pattern recognition methods, rely on the extraction of structures (*i.e.,* edges and curves) from the images using primitives, such as Chain Code Histogram (CCH). A limitation in these methods is that the images should be binary with defined boundaries, which can be a challenge in a real scenario.

2.2.2 OCR errors

Despite the high quality of modern OCR systems, Singh (2013) indicates a probability that one in 20 words is incorrect, which represents at least 5% of errors, and considering historical documents, this rate can be higher. As demonstrated by Bazzo et al. (2020), errors rate higher than 5% has a significant impact on retrieval quality.

The work by (NGUYEN et al., 2019) presents a study on the types of errors found in the text extraction of PDF documents. The authors also calculated the frequency of these errors using a set of monographs and newspapers produced between 1744 and 1921 in English and French. The words extracted incorrectly can be classified into two types.

- **Non-Words:** are words that are not in the lexicon of words considered correct, for example "*oil*"→"*oll*".
- **Real words:** are extraction errors that end up generating a word that appears in the lexicon of correct words, for example, "*week*"→"*weak*", which is a valid word.

Estimates by Nguyen et al. (2019) report that about 60% of errors are from real words while 40% are from non-words. Correcting errors that generate real words are more complex as they require that the context (*i.e.,* neighboring words in the sentence) are evaluated and this makes the method more costly from a computational point of view. The correction of non-words is also not trivial because it is difficult for a lexicon to contain all the possible correct words in a language – proper nouns, acronyms, several verb conjugations, and foreign words make this list incredibly large.

Segmentation errors are also frequent – they can be classified into two types.

- **Incorrect Segmentation:** occurs when a word is separated into two (or more), *i.e.,* an unexpected space is inserted into the word, for example "*number*"→"*nu mber*" or "*validate*"→"*valid ate*".
- **Incorrect Concatenation:** occurs when two (or more) words are concatenated into one, *i.e.,* the space is omitted, for example "*show image*"→"*showimage*" or "*in correct*"→"*incorrect*".

Segmentation problems are orthogonal to non-word and real word classification. Note that *valid*, *ate* and *in correct* are correct words whereas *nu*, *mber* and *showimage* are not, and would not be in the lexicon. Estimates by (NGUYEN et al., 2019) report that incorrect segmentation is 2.3 times more frequent than incorrect concatenation and that the two types of errors do not usually occur together.

2.2.3 Post OCR methods

Post-OCR methods aim to minimize the errors previously described, in Section 2.2.2, which is a challenging task. The errors associated with post-OCR text correction include not correcting a word that was incorrectly extracted and inserting errors in a word that was correctly extracted. In the last few years, the ICDAR (CHIRON et al., 2017; RIGAUD et al., 2019) organized two competitions for post-OCR text correction. The best-performing approaches employ state-of-the-art methods, such as deep learning algorithms (bidirectional LSTMs) using BERT (DEVLIN et al., 2018) embeddings as input. A recent survey by Nguyen et al. (2021) reports on the most recent advances and calls for approaches that address languages other than English.

Nguyen et al. (2021) divide these approaches into two main groups, manual and semi-automatic. Manual approaches consist mainly in collaborative systems to correct the documents, which, for large collections, are unfeasible. The semi-automatic approaches can be further divided into isolated-word and context-dependent techniques. Isolated-word correction can involve:

- merging OCR outputs (*i.e.,* combining outputs from different text extraction tools or different versions of the document);
- lexical approaches (*i.e.,* based on lexicons of word unigrams and string distance metrics);
- error models (*i.e.,* relies on models that represent common errors); and
- topic-based models,, which combine error and word unigram models.

The context-dependent approaches typically use language models (statistical or neural network-based) to take neighboring words into consideration. This way, they can fix both non-words and real-word OCR errors.

2.2.4 OCR Evaluation Metrics

To assess the quality of OCR extraction process, two error metrics are commonly used: *Character Error Rate* (CER) and *Word Error Rate* (WER) (CARRASCO, 2014). CER counts the number of character level operations that are required to transform the output into the ground truth and it is calculated as in Eq. 2.5. WER (Eq. 2.6) applies the same idea, but for word-level operations.

$$CER = \frac{i_c + s_c + d_c}{n_c} \tag{2.5}$$

$$WER = \frac{i_w + s_w + d_w}{n_w} \tag{2.6}$$

where i, s, and d represent insertion, substitution, and deletion, respectively. n_c and n_w are the number of characters and words in the ground truth.

2.3 Summary

In this chapter, we introduced fundamental concepts that will be approached along this work, which are related to OCR extraction and Information Retrieval. Here we presented the requirements for the creation of a test collection that represent a true IR setting to enable the proper evaluation of IR systems, such as the pooling method used to select candidate documents. In relation to OCR extraction, we introduced the main approaches used by OCR methods, the different type of errors found in OCR-ed documents, and statistics reported by researchers on this topic. Based on these concepts, the next chapter will discuss related work on the topics presented.

3 RELATED WORK

Over the years, many works have been done to evaluate different aspects that impact retrieval quality. More specifically, some authors were concerned about the problems caused by the OCR process. To evaluate IR systems, a test collection must have a set of components and requirements. In this chapter, we provide a general overview of the related works on the process of test collection creation and on evaluating the impact of OCR extraction and correction on IR.

3.1 Test collections

Over the years, many test collections for ad-hoc retrieval were created. Most of this effort was carried out within evaluation campaigns, such as TREC and CLEF. CLEF focused on European languages and thus, at the beginning of the 2000s, test collections were created for English, Italian, Spanish, Portuguese, German, French, Dutch, Finnish, Russian, and Swedish. The Portuguese collection, known as CHAVE (SANTOS; ROCHA, 2004), contains news articles published by Folha de São Paulo (Brazil) and Público (Portugal), 100 queries, and their relevance judgments. To the best of our knowledge, to this date, CHAVE is the only test collection for ad-hoc retrieval in Portuguese. This represents a serious limitation for the advancement of IR research in that language.

More recently, domain-specific test collections were also created for IR tasks. Lykke et al. (2010) constructed a test collection with documents on physics (monographies, papers, articles, and abstracts) with the purpose of evaluating integrated search. The collection contains 65 query topics and graded relevance assessments in four levels. Ritchie, Teufel and Robertson (2006) created a test collection with scientific papers from the ACL anthology. The goal was to explore evidence coming from the text of the citations (in analogy to the anchor text in web search). The collection has 170 queries, 7K documents, and graded relevance assessments in four levels. More recently, Basu et al. (2017) published a collection with microblog posts on disaster situations. The collection has about 50K tweets, five query topics, and binary relevance judgments. While the three aforementioned collections were in English, Soboroff, Griffitt and Strassel (2016) created BOLT, a multilingual passage retrieval collection. The documents consist of informal texts from discussion forums in English, Mandarin, and Arabic. The collection has 150 topics and over 2 million forum threads.

In the last few years, some efforts have been devoted to developing linguistic resources that can aid IR systems. Part of these resources was in the O&G domain in Portuguese and include corpora (CORDEIRO; VILLALOBOS, 2020), word embeddings (GOMES et al., 2021), and named-entity recognition systems (CONSOLI et al., 2020). While these resources can be useful to improve IR systems, there are no test collections available to assess them. In this work, our aim is to bridge this gap with REGIS, a Portuguese collection with relevance judgments of geoscientific documents.

3.2 Impact of OCR Extraction and Correction in Information Retrieval

The impact of OCR errors has been studied on a variety of tasks including contextual embeddings (JIANG et al., 2021), named entity recognition (MILLER et al., 2000; DUTTA; GUPTA, 2022; HAMDI et al., 2020; HUYNH; HAMDI; DOUCET, 2020; HEGGHAMMER, 2021; GUPTE et al., 2021), entity linking (PONTES et al., 2019), part-of-speech tagging (LIN, 2003), text summarization (JING; LOPRESTI; SHIH, 2003), text classification (ZU et al., 2004; HEGGHAMMER, 2021), and topic modeling (MUTUVI et al., 2018; HEGGHAMMER, 2021).

Specifically for IR, the pioneer studies that aimed at assessing the impact of OCR-ed text date back to the 1990's (TAGHVA et al., 1994; TAGHVA; BORSACK; CONDIT, 1996a; TAGHVA; BORSACK; CONDIT, 1996b; CROFT et al., 1994; WIEDENHOFER; HEIN; DENGEL, 1995). Taghva et al. (1994) 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 (45 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. Croft et al. (1994), 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%.

Mittendorf and Schäuble (2000) 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. Evershed and Fitch (2014), on the other hand, found an improvement of almost 60% in recall misses, *i.e.,* the number of unique ground-truth words in the corrected text was 60% higher than the number found in the uncorrected version. They worked with three datasets of historical texts, in three versions – extracted, ground truth, and automatically corrected. Along the same lines, Traub et al. (2018) studied the impacts of correcting OCR errors on document retrieval. Their dataset consisted of 100 issues of historic newspapers in Dutch for which they had ground truth and OCR-ed versions. Since there were no relevance judgments available, their analysis focused on the *retrievability score* which determines how often a document occurs when inspecting the top-k results for a set of queries. They found that high error rates correlate with low retrievability scores and that error correction leads to higher retrievability. In a similar fashion, Strien et al. (2020) 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 the quality of the documents decreases. More recently, (ZHUANG; ZUCCON, 2021) assessed the impact of typos in passage retrieval in dense vector representations. Their experiments with synthetic error insertion in 50% of the queries generated a loss of 24% in recall. The authors were able to mitigate this negative impact to 8% with typos-aware training in which at training time, the model sees both queries with and without typos.

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 (CROFT et al., 1994; BAZZO et al., 2020); and (ii) works using real OCR-ed documents, which lacked query topics and/or relevance judgements (EVERSHED; FITCH, 2014; TRAUB et al., 2018; STRIEN et al., 2020). To the best of our knowledge, only three works experimented with real OCR errors in a true IR setting (TAGHVA; BORSACK; CONDIT, 1996a; LAM-ADESINA; JONES, 2006; GHOSH et al., 2016).

Taghva, Borsack and Condit (1996a) worked with a small collection of 674 long documents and 59 query topics in English for which relevance judgments were produced. The authors found no significant differences in terms of mean average precision among

the ground-truth, automatically extracted, and corrected versions of the documents. Thus the main finding was that retrieval quality is not impacted by OCR issues. Lam-Adesina and Jones (2006) took an interesting approach to assemble their data collection – they took the TREC-8 spoken document retrieval collection (JOHNSON et al., 1999), 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. More recently, Ghosh et al. (2016) worked with documents from the FIRE RISOT collection in Bangla (68K documents and 66 query topics) and Hindi (94k 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.

The improvements found by Ghosh et al. (2016) and the lack of difference found by other research on English data (CROFT et al., 1994; TAGHVA et al., 1994; TAGHVA; BORSACK; CONDIT, 1996a; LAM-ADESINA; JONES, 2006), may suggest that the behavior varies across languages. A recent survey by Nguyen et al. (2021) 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. The language we use in this work, Portuguese, is not covered by existing OCR datasets.

Table 3.1 presents an overview of the related work compared to ours. We assess whether each publication dealt with real OCR errors (*i.e.,* if they used real PDF documents), if they had the corresponding ground truth (*i.e.,* text without errors), followed the standard IR evaluation procedure with queries and relevance judgments, the length of the documents used (where “L” corresponds to long, and “S” to short), whether they addressed OCR error correction, and which languages were used in the experiments. It can be seen that most works that conducted their experiments with real PDF documents did not use queries and relevance judgments (TAGHVA et al., 1994; WIEDENHOFER; HEIN; DENGEL, 1995; EVERSHED; FITCH, 2014; TRAUB et al., 2018; STRIEN et al., 2020). On the other hand, most works that followed the standard IR experimental procedure (CROFT et al., 1994; KANTOR; VOORHEES, 2000; MITTENDORF; SCHÄUBLE, 2000; BAZZO et al., 2020; VARGAS et al., 2021; ZHUANG; ZUCCON, 2021)

relied on synthetic OCR errors. Reinforcing the considerations made by (NGUYEN et al., 2021) and (BENDER, 2019), the majority of these works (10 out of 15) used only collections in English. Regarding document length, most works (ten out of 15) used only short documents in their experiments.

Table 3.1 – Comparison of related works that evaluate the impact of OCR errors. * indicates that the ground truth does not cover the entire set of documents

Work	Real PDFs	Ground Truth	Doc. Length	Relevance Judgments	Error Correction	Language
Croft et al. (1994)	✗	✓	S, L	✓	✗	EN
Taghva et al. (1994)	✓	✓	S, L	✗	✓	EN
Wiedenhofer, Hein and Dengel (1995)	✓	✓	S	✗	✓	DE
Taghva, Borsack and Condit (1996b)	✓	✓	L	✓	✓	EN
Taghva, Borsack and Condit (1996a)	✓	✓	L	✓	✓	EN
Kantor and Voorhees (2000)	✗	✓	S	✓	✓	EN
Mittendorf and Schäuble (2000)	✗	✓	S	✓	✓	EN
Lam-Adesina and Jones (2006)	✓	✓	S	✓	✗	EN
Evershed and Fitch (2014)	✓	✓	S	✗	✓	EN
Traub et al. (2018)	✓	✓	S	✗	✓	NL
Ghosh et al. (2016)	✓	✓*	S, L	✓	✗	EN, BN, HI
Strien et al. (2020)	✓	✓	S	✗	✗	EN
Bazzo et al. (2020)	✗	✓	S	✓	✗	PT
Vargas et al. (2021)	✗	✓	S	✓	✓	PT
Zhuang and Zuccon (2021)	✗	✓	S	✓	✗	EN
Ours	✓	✓*	S, L	✓	✓	PT

In this work, we build upon our previous work (BAZZO et al., 2020; VARGAS et al., 2021). We move to a realistic scenario by working with real extraction errors in a different IR collection. In comparison to the existing work presented in Table 3.1, our evaluation has the complete set of elements that enable assessing the impact of OCR extraction and error correction in a traditional IR experimental setting.

3.3 Summary

In this chapter, we presented the related work with respect to the creation of test collections and the evaluation of the impact of OCR extraction and correction on the quality of the IR results. The findings of these works let the discussion open – some results pointed that the post-processing can improve IR quality (EVERSHED; FITCH, 2014; VARGAS et al., 2021), while others found the opposite (TAGHVA et al., 1994; CROFT et al., 1994; MITTENDORF; SCHÄUBLE, 2000). A gap that was not covered by these works is that only a few works conducted their experiments with real OCR errors

in a true IR setting (TAGHVA; BORSACK; CONDIT, 1996a; LAM-ADESINA; JONES, 2006; GHOSH et al., 2016). As shown in Table 3.1, most of the cases that experimented with real PDF documents did not use queries and relevant judgments, while others that had a true IR setting, used OCR synthetic errors. By creating a test collection described in the next chapter, we try to fill these gaps.

4 REGIS COLLECTION

To evaluate an IR system, a test collection must have three components: (i) a set of documents, (ii) a set of query topics, and (iii) a set of relevance judgments for the query-document pairs.

The REGIS collection was created following the widely accepted recommendations by Manning, Raghavan and Schütze (2008) and Sanderson (2010). The pooling method, which is widely accepted and used to build many TREC test collections, was adopted in the assembling of REGIS. While most ad-hoc IR test collections employ binary relevance judgments (*i.e.,* a document is judged either as relevant or not relevant) REGIS adopts four levels of relevance – "very relevant", "fairly relevant", "marginally relevant", and "not relevant". These levels are useful in cases where the document does not answer the query completely or only a small piece of the document is related to the topic. Once these options are available to the annotators, possible relevant documents are less likely to be lost and deemed as not relevant. Furthermore, if necessary, it is easy to map the four relevance classes into binary judgments.

Following the well-established and widely used pattern, the documents are made available in the XML format. Each file is named with its *docid* and contains only one document with the following fields.

- `docid`: Document unique identification.
- `filename`: Name of the original file.
- `filetype`: Type of the original file.
- `text`: Document contents, extracted from the original PDF file.

4.1 Data Collection and Text Extraction

Our data can be divided into two categories: (i) technical reports and (ii) theses and dissertations. Technical reports were collected from two sources: the Brazilian Petroleum Agency (ANP)[1] and a Brazilian Oil Company[2]. These documents contain technical, scientific, and managerial information in the O&G domain. The theses and dissertations on the geoscientific domain were collected from the digital library of the Brazilian

[1] ANP: <https://www.gov.br/anp/pt-br/>
[2] Petrobras: <http://publicacoes.petrobras.com.br/>

Institute of Information in Science and Technology (IBICT)[3]. To select the documents that belong to the desired domain, we used keywords such as *"geology"*, *"petroleum"*, *pre-salt*, and *"sedimentary basin"*. The documents were created over a long period of time, dating from 1957 to 2020.

The original documents were in PDF. To extract their contents, we used two software that also provide OCR, namely ABBYY FineReader[4], and Tika[5]. Then, duplicated documents were detected and removed. Finally, a total of 21,444 documents remained. REGIS documents are typically very long, with an average of 25.1k tokens per document. The vocabulary of the collection is also large, with around 7M tokens (before stemming). This is due to the use of technical terms, proper nouns, misspellings, and OCR extraction errors.

4.2 Topic Building

The goal of our topic-building process was to mimic real user needs in the geoscientific domain. Thus, we tried to cover a broad range of topics and to assure a mix between generic and specific queries, as well as easier and harder ones. In order to achieve these goals, we had the collaboration of domain specialists who played a fundamental role in topic creation. These specialists created 27 query topics and provided all the *descriptions* and *narratives*. Also, to reproduce real user needs from the domain, some query topics were taken from logs of real searches submitted to a retrieval system in a Brazilian oil company. The logs consisted simply of queries (*i.e.,* sets of keywords) and their submission time. The queries on the log were filtered to keep the ones that contained between three and ten tokens. Then, from these resulting queries, the specialists selected the most representative ones. The nine selected queries were transformed into topics composed of id, title, description, and narrative as follows:

- `num`: Topic unique identifier.
- `title`: A short title, which is commonly used as the query submitted to the IR system.
- `desc`: A short description of the information need, generally, with no more than one sentence.

[3]IBICT: <https://bdtd.ibict.br/vufind/>
[4]ABBYY: <https://www.abbyy.com/>
[5]Tika: <https://tika.apache.org/>

- `narr`: A more detailed narrative that helps the annotator decide on the relevance of the documents.

At the end of this process, there were 36 candidate queries. Those were uploaded into the annotation system. Figure 4.1 shows an example of a query topic.

Figure 4.1 – Example of a query topic in REGIS. The topic describes an information need to find seismic data from the Sergipe-Alagoas Basin in articles, theses, dissertations, monographs, or reports, which preferably mention correlation among wells.

```
<top>
    <num> 28 < \num>
    <title> Sísmica de Sergipe-Alagoas.< \title>
    <desc> Buscar por documentos que abordem dados
sísmicos da Bacia Sergipe-Alagoas.< \desc>
    <narr> Documentos de interesse incluem artigos, teses,
dissertações, monografias ou relatórios, que tenham como
tema específico dados sobre sísmica e de preferência
correlação entre poços da Bacia Sergipe-Alagoas.< \narr>
< \top>
```

Source: The Authors

4.3 Pool Creation

Building a test collection in which annotators judge the query-document pairs is unfeasible. To address this problem, we adopted the pooling methodology, already described in Section 2.1.1. This method was proposed by Spark-Jones (1975) and became the standard procedure for the creation of test collections.

As the recommendation is to use more than one IR system and/or different ranking functions, our pool was created using two IR systems, Apache Solr[6] and Anserini[7]. In order to select the best configurations, some preliminary experiments were run with the CHAVE (SANTOS; ROCHA, 2004) test collection. The four configurations (two from each IR system) are presented in Table 4.1. Having different configurations is important to avoid system bias (SANDERSON, 2010). In Solr, we used Okapi BM25 and DFR (Divergence From Randomness) as scoring functions along with proximity search to give a greater score to documents in which the terms of the query are closer. In Anserini, we used BM25 combined with RM3 (Relevance Model 3) (*i.e.,* language modeling for query expansion), and QLD (Query Likelihood with Dirichlet smoothing). Stemming was

[6]Apache Solr: <https://lucene.apache.org/solr/>
[7]Anserini: <http://anserini.io/>

Table 4.1 – IR system configurations

Id	IR System	Scoring function	Search options
BM25+Prox	Solr	BM25	Proximity search
DFR+Prox	Solr	DFR	Proximity search
BM25+RM3	Anserini	BM25	RM3
QLD	Anserini	QLD	–

Source: The Authors

applied in all runs – Lucene Portuguese Light Stem was used in Solr and, in Anserini, we applied the default Porter stemmer.

The keywords submitted to the search engines were created by simply taking the title field of the topics. We took the union of the top 50 documents from each of the four configurations. The four original rankings were aggregated by considering the number of rankings and the position in which the document appeared. Then, the resulting pool for each query was composed of the 50 candidates according to the aggregated ranking.

Then, in an effort for the pool to contain all relevant documents for a query, a specialist issued modified versions of the queries including synonyms and related words. Whenever a new potentially relevant document was found (*i.e.,* one that was not already in the pool for the query), it was added to the pool.

4.4 Relevance Assessments

To enable assessing the relevance of the documents with respect to the query topics, a complete annotation system was developed. This section describes the creation of the system and the entire annotation process.

4.4.1 Annotation System

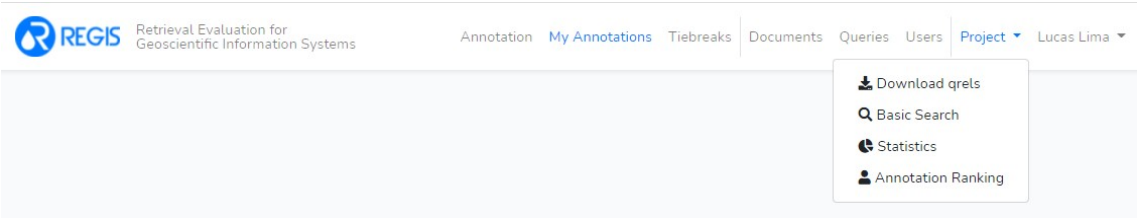
The annotation system was developed using Laravel[8], an open-source PHP framework, with a robust architecture that follows the MVC (Model-View-Controller) design pattern. The model layer is related to the data operations, logic, and rules, for which MySQL was the database adopted. The view layer corresponds to all visual elements and user interactions, for which the default web tools HTML, CSS, and JavaScript were used. Finally, the controller is responsible for the integration of the other layers, receiving

[8]Laravel: <https://laravel.com/docs/8.x>

requests and returning responses.

The REGIS annotation system[9] allows the entire CRUD (create, read, update, and delete) operations over queries, documents, and annotations. The interface has two user levels, *administrator* with full privileges, and *annotator*, with restricted access allowing only the processes related to the performing relevance judgments. The entire header of the system can be seen in Figure 4.2. Annotators had access to the first three menu items, which correspond respectively to the annotation page, the user's own annotations (allowing them to edit if necessary)the , and tiebreaks page (which lists all query-document pairs for which the annotators disagreed).

Figure 4.2 – Screenshot of the admin's header from REGIS system



Source: The Authors

The Administrator interface has control over all queries. Figure 4.3 presents the list of candidate topics, in which it is possible to follow the progress and manage each one of them, including their related documents. Figure 4.4 shows the system dashboard, with general statistics of the annotation progress, the status of the query topics, relevance judgments, query distribution, and the ranking of annotators – a public ranking also was made available to try to motivate the annotators.

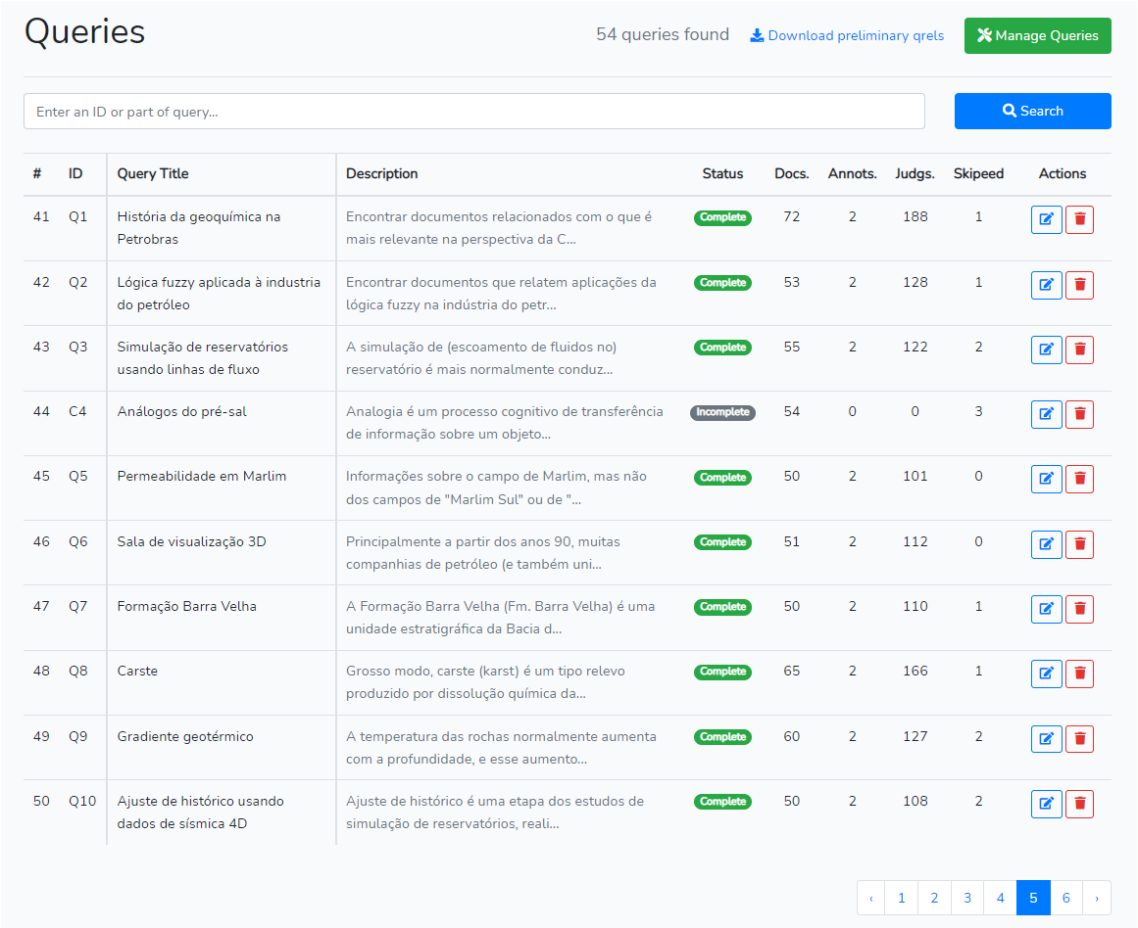
In the annotation page, the system presents the description of the information need, the document with the query terms highlighted, a link to the original PDF document, and the relevance classes. The annotator could also enter any comments they felt were important for the relevance assessment. In addition, if the annotator felt that the query fell outside their area of expertise, they could choose to skip the query. Figure 4.5 shows a screenshot of the annotation page.

4.4.2 Annotation Process

To assure the quality of the collection, the relevance judgments were made by annotators with domain knowledge which included geologists and petroleum engineers.

[9]REGIS system: <https://github.com/lucaslioli/regis-system>

Figure 4.3 – Screenshot of the queries page from the REGIS system



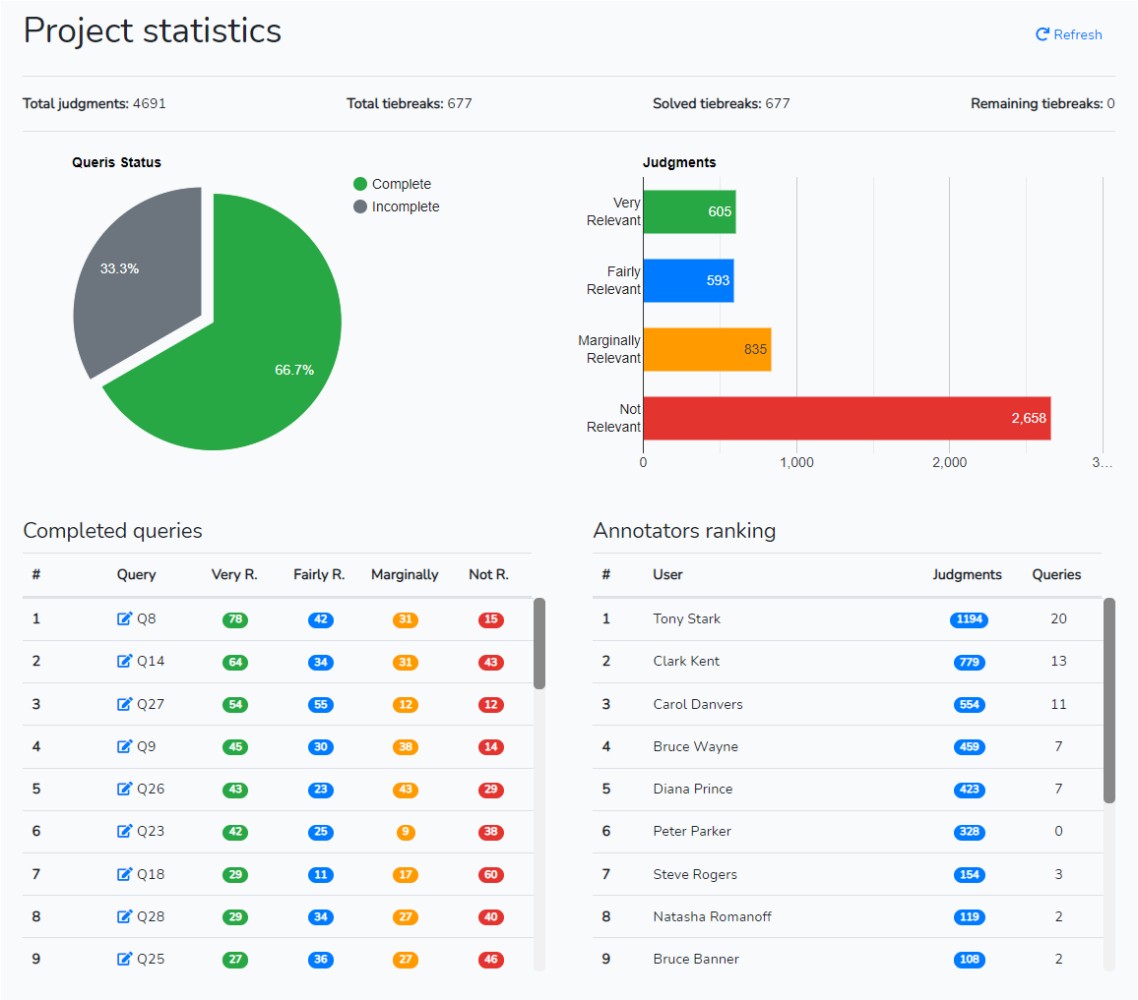
Source: The Authors

The subjects were recruited from the Geosciences department in a Brazilian university and from Petrobras, the main Brazilian oil company. Also, to increase the confidence in the judgments, each query-document pair was judged by at least two annotators. In cases where the annotators disagreed, a third annotator was summoned to break the tie. Documents were presented in order of doc-id (and not in the order returned by the scoring functions) to avoid ranking bias.

Our annotation effort was carried out by 16 assessors who made a total of 4691 judgments. These numbers include 667 tiebreaks. On average, the annotators spent around two and a half minutes assessing each document/query pair. Non-relevant documents were faster to judge while distinguishing among the levels of relevance tends to demand a more careful evaluation. We estimate that the overall time taken was around 230 hours. We calculated the inter-annotator agreement according to Fleiss kappa. The obtained score was 0.392, which shows fair agreement.

From the 36 queries that were judged, two did not have any documents considered at least *fairly relevant* and were discarded. Thus, at the end of the process, REGIS has

Figure 4.4 – Screenshot of the statistics page from REGIS system



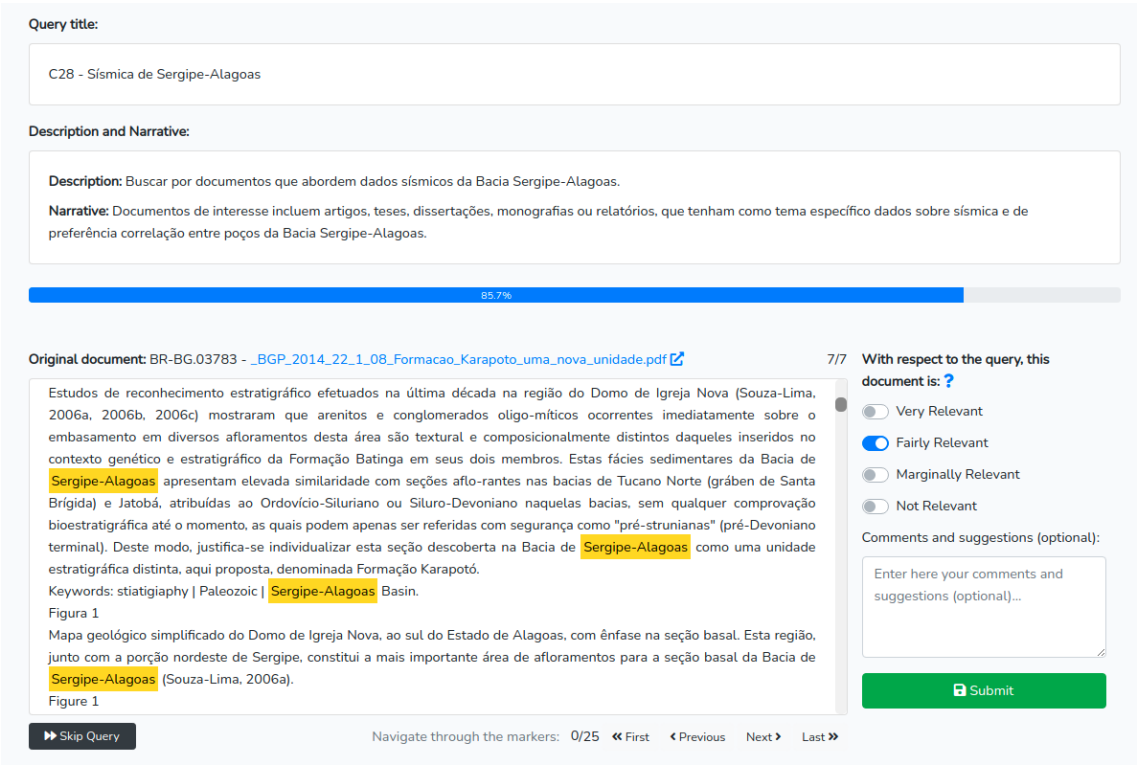
Source: The Authors

34 queries. While smaller than the number of queries normally found in generic-domain test collections, this number can be considered enough to allow experimenting with retrieval techniques. In an experimental evaluation of several evaluation metrics, Buckley and Voorhees (2017) found that, for mean average precision, 25 topics are the minimum number considered acceptable.

Figure 4.6 shows the distribution of the levels of relevance of the judged documents by query. We can see a wide variation ranging from queries that have all judged documents being rated as at least marginally relevant (Q8) to queries in which no document was classified as very relevant (Q4, Q11, Q15, and Q34). We believe that this shows we accomplished the goal of assuring queries with different levels of difficulty.

Based on preliminary experiments, Figure 4.7 shows the average precision results for the individual queries on BM25+Prox. We found a moderate Pearson correlation of 0.47 between the number of relevant documents and average precision. This indicates a tendency of queries with more relevant documents yielding better results.

Figure 4.5 – Screenshot of the annotation page of the system



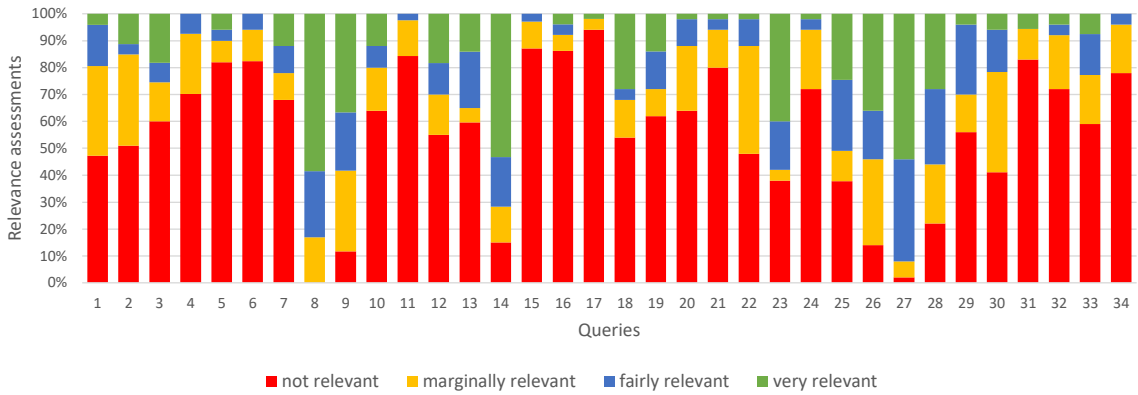
Source: The Authors

4.5 Discussion

In this section, we discuss some relevant issues and limitations related to the construction of REGIS and general test collections.

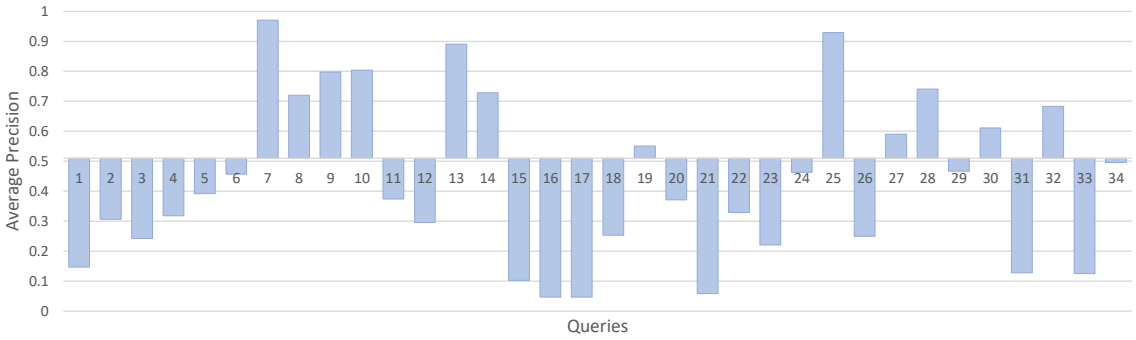
Annotation quality. Having good quality relevance assessments is key for a test collection. We tried to ensure a high annotation by following best practice guidelines in the area. Analyzing the disagreements between judges, we found that most of them (71%) involved relevance levels that were immediately above/below one another. Disagreements

Figure 4.6 – Distribution of the levels of relevance



Source: The Authors

Figure 4.7 – Average Precision by Query for BM25 with proximity under Solr



Source: The Authors

between highly relevant and not relevant assessments accounted for only 2% of the cases. Having a third annotator to solve the disagreements was important to ensure consistency. As pointed by Sanderson (2010), some studies have already been done about the assessor consistency, and their experiments showed that variations in judgments do not have a high impact on the ranking, considering different configurations of an IR system.

Document Length. REGIS has long documents and this represents both advantages and disadvantages in terms of retrieval quality. Long documents can contain all words in the query (possibly many times) and still not be relevant as these words could be far apart or mentioned in different contexts. Thus the importance of proximity search.

Text Extraction Errors. The original source files of some documents in REGIS are scanned images from the physical document. Thus, we had to resort to OCR software to extract the textual contents. During the annotation process, extraction errors became evident. The work by Bazzo et al. (2020) has shown that if such errors exceed a 5% word error rate, then retrieval quality can be significantly affected. As the scanned documents are not the majority, we believe these errors fall below 5% word error rate. But, considering only this scanned portion, which contains historical documents, this rate can easily increase.

Syntax difference. Some important words in the geoscientific domain can have different spellings such as *pré-sal* and *pré sal* or *carste* and *karste*. This issue is yet more present in REGIS as the documents were written over a long period of time (over 60 years), during which spelling reforms took place and changed the orthography of several words. In addition, despite being in Portuguese, several technical terms can be used in English as well. As a result of these syntax issues, having good results for some queries may be quite challenging.

Limitations. Finally, there are several criticisms of the traditional IR evaluation paradigm based on relevance judgments. Some experimental evaluations identified that

the results of batch and user searching could be different (HERSH et al., 2000; TURPIN; SCHOLER, 2006). The traditional paradigm cannot assess all elements that are important in a search experience. Nevertheless, test collections still are valuable resources that can yield a series of insights on how to improve retrieval quality.

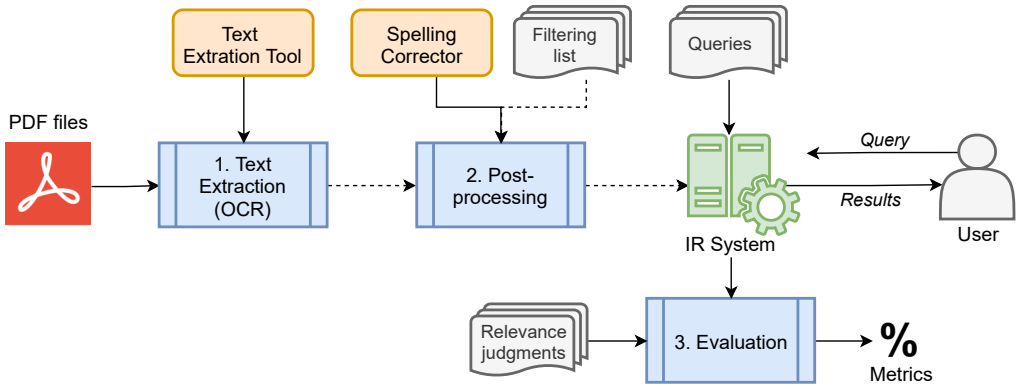
4.6 Summary

In this chapter, we described the entire creation process of the REGIS, test collection, including the document collection, text extraction, topic creation, and the generation of relevance assessments. The development of the collection required the creation of a complete annotation system to support the process. REGIS is a collection for the geoscientific domain in Portuguese, which contains 20K documents and 34 query topics along with relevance assessments. The documents are typically very long, with an average of 25.1k tokens per document, and the preliminary experiments showed a good distribution of the queries difficulty. This difference in difficulty levels is essential to evaluate the quality of IR systems.

5 MATERIALS AND METHODS

To assess the impact of OCR extraction and post-processing on IR quality we follow the process depicted in Figure 5.1. We start by taking the original PDF documents and extracting their textual contents (1). Once the text is extracted, it may be submitted to a post-processing step that aims at fixing extraction errors (2). Then, at the Evaluation stage (3), the queries are run and scored using the relevance judgments.

Figure 5.1 – Evaluating the impact of text extraction and correction on Information Retrieval



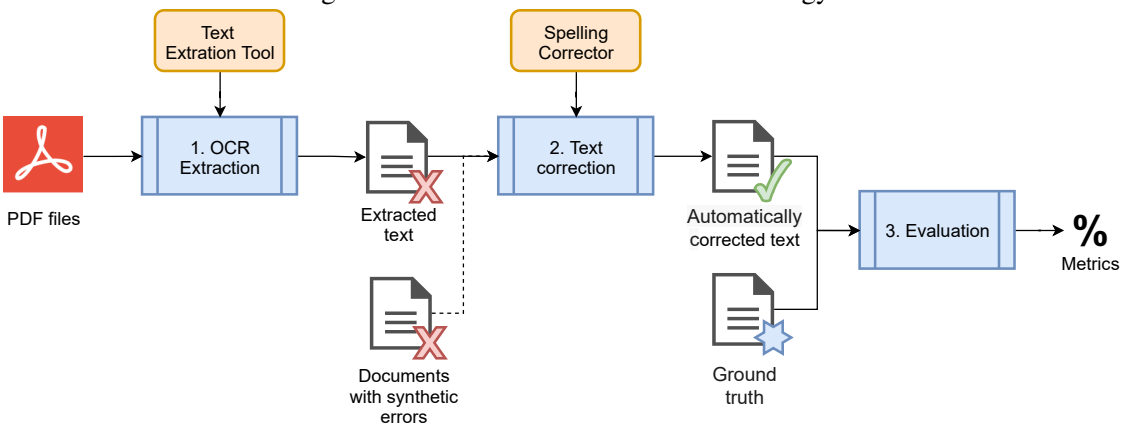
Source: The Authors

The pipeline described in Figure 5.1 can be seen as an *extrinsic evaluation* since it assesses how extraction/correction impacts a downstream task. Nevertheless, it is also important to have an *intrinsic* evaluation to provide insights about the inherent quality of the text extraction and correction processes, regardless of the retrieval results. The intrinsic evaluation pipeline is described in Figure 5.2 and it requires a ground truth. The main phases for these experiments are similar to the ones in Figure 5.1. The difference is in the evaluation step, which uses the ground truth of text extraction 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.

5.1 Test Collections

The ideal test collection for our experimental setting would require four components to enable a complete evaluation of the quality of the text extraction 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 (CHIRON et al., 2017; RIGAUD et

Figure 5.2 – Intrinsic evaluation methodology



Source: The Authors

al., 2019) are not suitable because they lack queries and relevance judgments (*i.e.,* components iii and iv). IR test collections typically lack component i, as they work with text that was originally in a digital format.

In addition to the REGIS collection, to complement the experiments and compare the results, the other test collection used during the experiments was **CHAVE** (SANTOS; ROCHA, 2004), which is composed of news articles from the Brazilian newspaper Folha de São Paulo, dating between 1994 and 1995. The structure in this collection, such as query topics and documents, is very similar to REGIS, which facilitates performing the experiments. Details about the documents and tokens from both collections are shown in Table 5.1. Appendix A shows examples of documents from both collections. Since the documents from REGIS can be very long, only the abstract were shown.

Because CHAVE is not an OCR-ed collection, it was necessary to adopt the alternative version produced in Bazzo et al. (2020), which inserted along with the documents 25% of synthetic OCR errors. The insertion followed probabilities distribution for OCR errors observed in real cases, considering character exchanges, inserting spaces into words, removing spaces between words, and inserting erroneous symbols.

Table 5.1 – REGIS and CHAVE Statistics

Statistic	**REGIS**	**CHAVE**
Documents	21,444	103,913
Tokens	538.4M	17.5M
Queries	34	100
Distinct Tokens	4.1M	655.8K
Avg Tokens per doc	25.1K	168

Source: The Authors

5.2 OCR Tools

The original documents from REGIS were in PDF and about 40% of them were not digitally created, so OCR tools were needed to extract their textual contents. Three solutions were used in this task:

- **Apache Tika**[1], an open-source application that can be used through the command line to parse and extract information from many different types of file. It uses the Tesseract Parser[2] as the OCR mechanism, and a PDF Parser bases on heuristics to decide when to run OCR over a document page.
- **ABBYY FineReader**[3], a desktop software that provides many features to manipulate and process PDF files, such as digitize, retrieve, edit, etc. ABBYY uses AI-based OCR methods, and as a proprietary software, there is not much information available about the inter processes. In our experiments we used version 14 of this tool.
- **Tornado**[4], 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[5], Detectron2[6], PDFMiner[7], Camelot[8], Tesseract OCR Parser, and Luigi[9]. 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 extraction in the Oil and Gas industry, which includes the geoscientific domain of the REGIS collection. Due to the cost, this method was run using resources from the Brazilian supercomputer SDumont[10].

5.3 Post-processing Methods

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

[1]Apache Tika: <https://tika.apache.org/>
[2]Tesseract OCR Parser: <https://github.com/tesseract-ocr/tesseract>
[3]ABBYY FineReader: <https://pdf.abbyy.com/>
[4]Tornado: <https://petroles.puc-rio.ai>
[5]Poppler: <https://github.com/freedesktop/poppler>
[6]Detectron2: <https://github.com/facebookresearch/detectron2>
[7]PDFMiner: <https://github.com/pdfminer/pdfminer.six>
[8]Camelot: <https://github.com/camelot-dev/camelot>
[9]Luigi: <https://github.com/spotify/luigi>
[10]SDumont: <http://sdumont.lncc.br/>

- SymSpell[11], 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. As input for this method we provided the dictionaries based on abstracts from a portion the documents in REGIS, aiming to include technical vocabulary and some named entities from the specific domain.
- `sOCRates` (VARGAS et al., 2021) 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. For the potentially wrong words, the method selects the most appropriate correction based on candidates obtained by two other methods, ASpell[12] and also SymSpell.

Post-processing can be time-consuming. Thus, aiming to improve efficiency, a filtering list was used to avoid post-processing the digitally created PDF documents. This way, the documents that have not been scanned did not go through the correction step. The list was obtained by analyzing the metadata of the files, more specifically, the creation and production tools.

5.4 IR Systems

The IR system used in our experiments was Apache Solr, which is an open-source search platform based on Apache Lucene. The configurations adopted were the same from the best result obtained to create REGIS 4, with BM25 as the scoring function and Portuguese Light Stemmer. 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.

[11] SymSpell: <https://github.com/wolfgarbe/SymSpell>
[12] <http://aspell.net/>

5.5 Evaluation Metrics

The evaluation metrics used in the extrinsic experiments are the standard IR metrics described in Section 2.1.2, namely MAP, and NDCG. Our analysis also focused on the raw number of relevant retrieved (*Rel. Ret.*). All these metrics were computed using Trec_eval[13]. To assess whether the differences were statistically significant, we used T-tests with α = 0.05. 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 (BUCKLEY; VOORHEES, 2017).

Since the relevance assessments in REGIS are graded in 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.

For the intrinsic experiments, we calculated the error metrics CER and WER (described in Section 2.2.4) using the OCREvaluation script[14].

5.6 Experimental Runs

Due to the high cost to process huge amounts of documents and time spent by some methods, we limited the execution of the post-processing methods over only Tika version. Which resulted int the following five experimental runs:

1. **Tika (baseline)**: Text extracted by Tika - no post-processing;
2. **ABBYY**: Texts extracted by ABBYY - no post-processing;
3. **Tornado**: Text extracted by Tornado - no post-processing.
4. **Tika + `sOCRates`**: Text extracted by Tika and post-processed by `sOCRates`;
5. **Tika + SymSpell**: Text extracted by Tika and post-processed by SymSpell;

[13]Trec_eval: <https://trec.nist.gov/trec_eval/>
[14]<https://github.com/impactcentre/ocrevalUAtion>

5.7 Summary

In this chapter, we described the two methodologies adopted in the extrinsic and intrinsic experiments along with the materials and methods used throughout the process, and the evaluation metrics. Besides REGIS, to complement the experiments and compare the results, we also performed the experiments on the CHAVE collection (SANTOS; ROCHA, 2004). Three OCR tools were used to extract the text from the real PDF documents in REGIS, and two post-processing methods were used aiming to fix the extraction errors. The results of the experiments are shown in the next chapter.

6 EXPERIMENTAL RESULTS

Following the methodologies presented and using created collection REGIS, in this chapter we will present the experimental results, trying to answer our two main research questions: (i) Which OCR extraction system performs better? (ii) Can error correction improve retrieval results? Complementing that, we realized a topic-by-topic analysis, intrinsic evaluation using a manually created ground truth, and compared the results from REGIS with CHAVE collection.

6.1 The impact of OCR Quality on Retrieval Effectiveness Metrics

Table 6.1 shows the results for three text extraction 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 text extraction 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. 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.

These results show that text extraction quality indeed impacts retrieval evaluation metrics – even for long documents such as the ones in REGIS.

6.2 Impact of Error Correction on Retrieval Results

To answer this question, we took the text extracted by Tika and post-processed it with `sOCRates` and SymSpell. Tika was chosen because it had the lowest scores in Table 6.1, which means it has more room for improvement. The results of this analysis are in Table 6.2. Looking at lines 2, 3, 5, and 6 we can see that `sOCRates` was better than SymSpell in terms of MAP, relevant retrieved, NDCG in the tolerant scenario and, in the strict scenario, in terms of MAP and NDCG, although both methods obtained the

Table 6.1 – Information retrieval quality metrics for the OCR extraction systems. The tolerant scenario considers "Marginally Relevant" and the strict scenario considers "Fairly Relevant" as the minimum relevance level. Best results in bold.

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

Source: The Authors

same number of relevant retrieved. On the other hand, considering efficiency, SymSpell works approximately nine times faster than `sOCRates`.

However, for both scenarios, the correction tools resulted in worse performance compared to the Tika run but 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 Section 6.3 we perform a topic-by-topic analysis.

Table 6.2 – Results for OCR error correction. The tolerant scenario considers "Marginally Relevant" and the strict scenario considers "Fairly Relevant" as the minimum relevance level.

Configuration	MAP	Δ	Rel.Ret	Δ	NDCG	Δ
Tolerant Scenario						
(1) Tika	**.4947**	–	**657**	–	**.6705**	–
(2) + `sOCRates`	.4904	-0.87%	652	-0.76%	.6664	-0.61%
(3) + SymSpell	.4810	-2.77%	648	-1.37%	.6566	-2.07%
Strict Scenario						
(4) Tika	**.4549**	–	**420**	–	**.6705**	–
(5) + `sOCRates`	.4428	-2.66%	417	-0.71%	.6664	-0.61%
(6) + SymSpell	.4313	-5.19%	417	-0.71%	.6566	-2.07%

Source: The Authors

6.3 Topic-by-Topic Analysis

Figure 6.1 presents MAP results by topic for the three extractors and for the two post-processed versions, each one represented by a different color. For each query topic, the MAP scores are sorted in decreasing order from left to right. The idea is to help understand how each configuration performs at each topic. As presented early in Chapter 4, there is a mix between easier and harder query topics in the REGIS collection. 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 MAPs were achieved by ABBYY, which can be seen by the large concentration of orange cells in the first column (21 out of 34). Despite `sOCRates` being better than SymSpell on the average of the whole set of query topics, Tika + SymSpell appears six times in the first column while Tika + `sOCRates` is never the first.

Over half the query topics (19 out of 34) were positively influenced by text correction with SymSpell (Q3-Q5, Q15-Q21, Q23, Q26, Q28-Q34). These are the cases in Figure 6.1 in which the green cells (Tika + SymSpell) ranked higher than the red cells (Tika). Queries Q6 and Q10 were the ones in which text correction with SymSpell had a very negative impact (decreases of 80 and 52%, respectively). `sOCRates`, on the other hand, only improved six topics (Q5, Q6, Q9, Q25, Q27, Q28) but with very small changes.

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 were observed in Q17, for which 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.

Following the pattern of visual resources used in other works to evaluate this type of data (FLORES; MOREIRA, 2016; BUCKLEY; VOORHEES, 2017), Table 6.3 com-

Figure 6.1 – MAP results of each configuration sorted in decreasing order for all query topics in the tolerant scenario.

Topic	1st	2nd	3rd	4th	5th
Q1	0.2487	0.2428	0.2417	0.2061	0.1936
Q2	0.4434	0.4417	0.4331	0.4289	0.4164
Q3	0.1804	0.1785	0.1759	0.1613	0.1597
Q4	0.4775	0.4641	0.4549	0.4524	0.4414
Q5	0.6051	0.5779	0.5468	0.5170	0.5020
Q6	0.4645	0.4463	0.4456	0.3746	0.0900
Q7	0.9612	0.9612	0.9592	0.9478	0.9325
Q8	0.7799	0.7639	0.7631	0.7486	0.7289
Q9	0.8941	0.7557	0.7496	0.7492	0.7386
Q10	0.7482	0.7332	0.6908	0.6868	0.3285
Q11	0.3884	0.3788	0.3734	0.3642	0.3273
Q12	0.4656	0.4355	0.4283	0.4167	0.4036
Q13	0.9006	0.8483	0.7577	0.7577	0.7448
Q14	0.8344	0.7622	0.7175	0.7094	0.7028
Q15	0.3658	0.3330	0.3330	0.3034	0.2835
Q16	0.1104	0.1036	0.0829	0.0829	0.0815
Q17	0.1111	0.1111	0.0667	0.0667	0.0667
Q18	0.2230	0.1977	0.1910	0.1909	0.1833
Q19	0.2721	0.2567	0.2547	0.2493	0.2482
Q20	0.6468	0.5645	0.5316	0.5212	0.4980
Q21	0.6568	0.5592	0.5500	0.5485	0.5365
Q22	0.8588	0.8526	0.8307	0.8304	0.8192
Q23	0.4819	0.4813	0.4551	0.4495	0.4494
Q24	0.6497	0.5768	0.5673	0.5505	0.5492
Q25	0.9459	0.9215	0.9167	0.9158	0.9140
Q26	0.6205	0.6137	0.6127	0.6072	0.5706
Q27	0.6806	0.6755	0.6717	0.6630	0.6592
Q28	0.6526	0.6396	0.5834	0.5631	0.5628
Q29	0.7157	0.6496	0.6221	0.5902	0.5902
Q30	0.7430	0.5992	0.5275	0.5127	0.5107
Q31	0.2517	0.2344	0.0870	0.0846	0.0846
Q32	0.7580	0.7235	0.7117	0.7018	0.5160
Q33	0.1904	0.1899	0.1861	0.1804	0.1731
Q34	0.5451	0.4633	0.4224	0.4156	0.4156
All	0.5438	0.5054	0.4947	0.4904	0.4810

Tika	ABBYY	Tornado
Tika + SymSpell	Tika + sOCRates	

Source: The Authors

Table 6.3 – Pairwise comparisons of all experimental runs. The cells show the number of topics in which the configuration of the column is better than (green), equivalent (blue), or worse than (red) the configuration of the row. Proportional differences consider a 5% margin.

	Tika + sOCRates			**Tika + SymSpell**			**ABBYY**			**Tornado**		
					MAP							
Tika	1	33	0	2	24	8	2	14	18	8	13	13
Tika + sOCRates				2	21	11	2	11	21	6	14	14
Tika + Symspell							5	14	15	10	12	12
Abbyy										19	10	5
					Rel.Ret							
Tika	1	33	0	4	30	0	3	18	13	7	19	8
Tika + sOCRates				4	29	1	3	17	14	6	20	8
Tika + Symspell							4	14	16	6	17	11
Abbyy										13	19	2

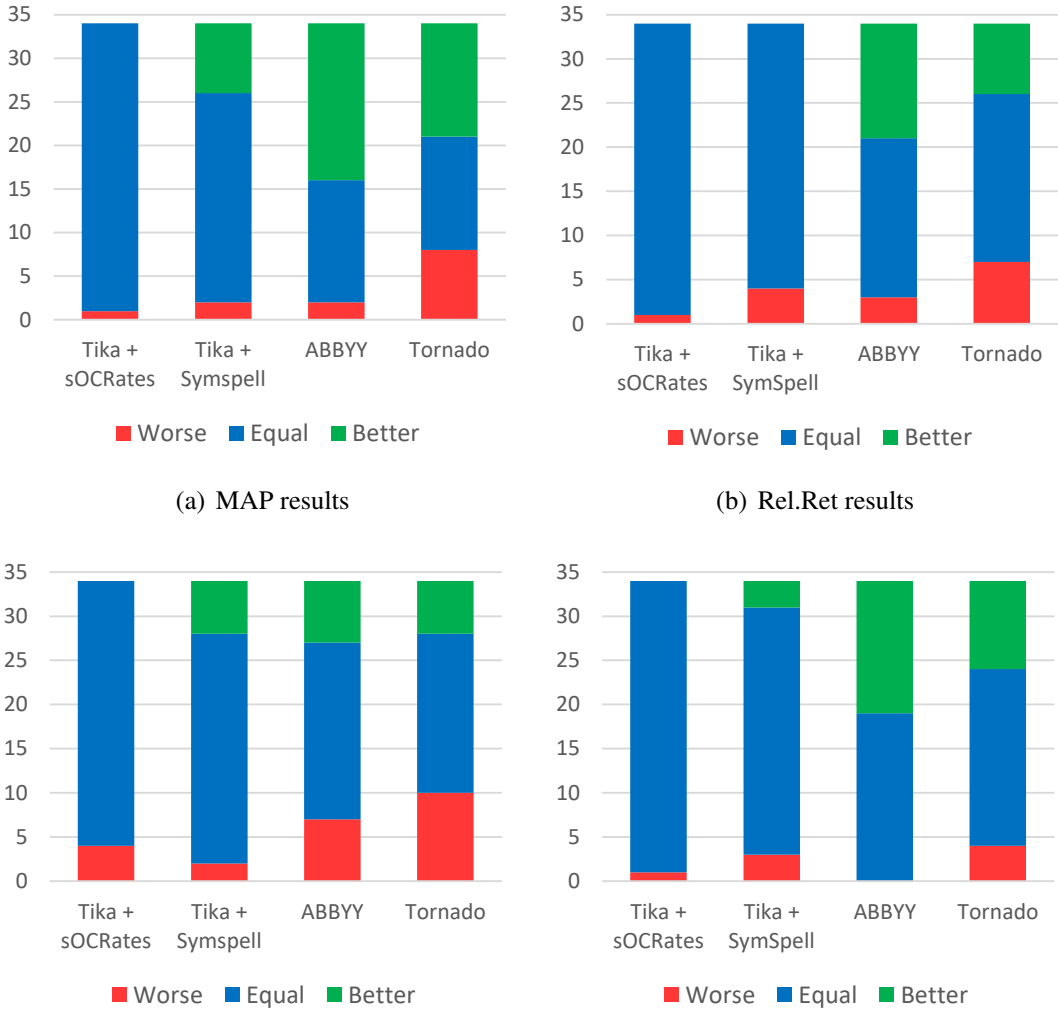
Source: The Authors

putes, for each pair of configurations, how many times the configuration on the column was worse, equivalent, or better than the configuration of the row. Results within a 5% proportional difference were considered equivalent (*i.e.,* Tornado was worse than ABBYY in 19 topics, equivalent in 10, and better in 5). Looking at the text extraction 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 reductions in two. `sOCRates`, on the other hand, had very small changes in comparison to the Tika baseline.

Figure 6.2 takes Tika as a baseline and compares the other runs considering a 5% margin. The results were distributed between worse, equal, or better. From these charts, we can observe that `sOCRates` did not present a great difference from the extracted version, and the post-processing methods did not bring improvements, considering the relevant documents retrieved. Observing Pr@10 metric, except `sOCRates`, the three other versions had similar improvements, although in the case of Tornado, the number of worse cases was bigger. As demonstrated in previous experiments in Section 6.2, the better results from ABBYY can also be seen here, this version obtained more better than worse cases in three of the four metrics, being that for NDCG there was no case of worsening and for PR@10 the number was the same.

Figure 6.3 shows the precision-recall curves for the tolerant and strict scenario. Both graphs reinforce what is shown in Table 6.2 and Figure 6.2, ABBYY had the best

Figure 6.2 – Number of topics in which each configuration obtained worse, equal, or better results in comparison to Tika (considering a 5% margin).



(a) MAP results

(b) Rel.Ret results

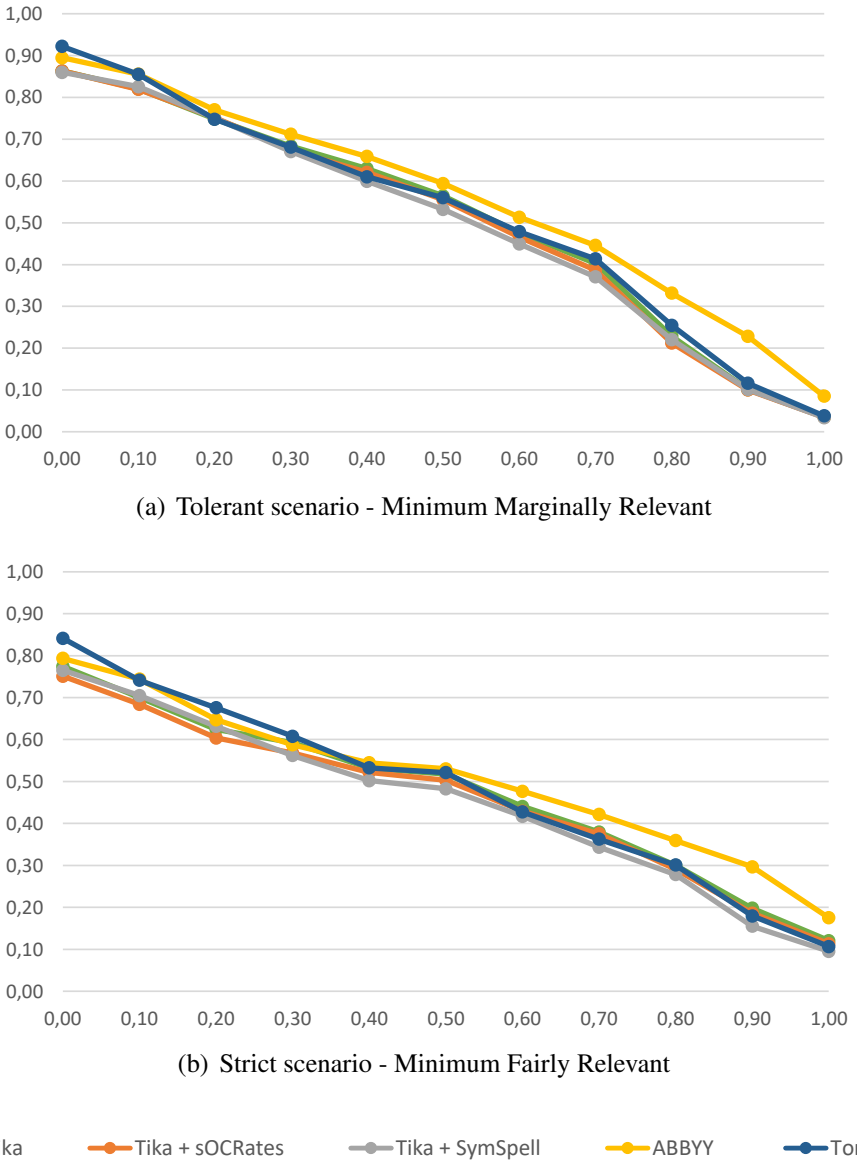
(c) PR@10 results

(d) NDCG results

Source: The Authors

performance, while other configurations obtained similar performances.

Figure 6.3 – Precision-recall curves for the tolerant scenario - Minimum Marginally Relevant



(a) Tolerant scenario - Minimum Marginally Relevant

(b) Strict scenario - Minimum Fairly Relevant

Source: The Authors

6.4 Intrinsic Evaluation

In our intrinsic evaluation, we compare the results of OCR extraction and correction against the ground truth of text extractions for a sample of sentences. Table 6.4 presents the results 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 text extraction tools –

ABBYY is the best performing, followed by Tornado, and Tika comes last.

In a related investigation, Hegghammer (2021) 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. We can see that our results are closer to the quality obtained for the English documents.

Regarding the correction methods, we see that in many cases they failed to fix problems with the Tika extractions 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. This confirms the findings by Vargas et al. (2021) 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.

Table 6.4 – Intrinsic results and Pearson correlation with retrieval quality metrics. CER and WER are error metrics, so the lower the better.

Method	**CER** ↓	**WER** ↓	**MAP** ↑	**NDCG** ↑	**Rel.Ret** ↑
Tika	1.09	6.49	0.4947	0.6705	657
Tika + `sOCRates`	2.16	7.25	0.4904	0.6664	652
Tika + SymSpell	3.78	11.84	0.4810	0.6566	648
ABBYY	**0.31**	**1.57**	0.5438	0.7109	697
Tornado	3.17	8.75	0.5054	0.6911	666
Pearson Correlation		**CER**	-0.7143	-0.5932	-0.6891
		WER	-0.8828	-0.7949	-0.8584

Source: The Authors

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 domains dictionaries with bigrams and unigrams are necessary.

In the work from Bazzo et al. (2020), in order to insert the synthetic errors into the CHAVE collection, the analysis performed to identify the pattern of OCR errors found that some common errors were 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. Complementing that analysis, in Table 6.5 we present some other examples of words extracted by Apache Tika and post-processed by SymSpell and `sOCRates`. Lines

1-7 show examples of erroneous extractions (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 extractions 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.

Table 6.5 – Examples of words extracted by Tika and their corresponding version post-processed. Correct words have a ✓and incorrect words have a ✗.

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) injeção ✓	indec a o ✗	injeção ✓
(10) aplicação de técnicas ✓	aplicar a o de tecnicas ✗	aplicação de técnicas ✓
(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 ✓

Source: The Authors

6.5 Comparing With Another IR Test Collection

In order to compare the impacts of OCR correction in another document collection, we reproduced our extrinsic experiments using the CHAVE test collection (SANTOS; ROCHA, 2004). While REGIS is composed of large domain-specific documents, CHAVE is composed of short newspaper articles. Since the input documents in CHAVE are already in pure text, we took the version with synthetically inserted errors created by Bazzo et al. (2020) and used by Vargas et al. (2021).

The same Solr configurations were used in both collections. The baseline run in REGIS is the one in which the text extraction system was Apache Tika, and the baseline

in CHAVE has synthetic errors inserted in 25% of the words. In CHAVE, we also have an *ideal* run with the original clean texts that works as an upper bound for the correction systems. Table 6.6 presents the results for this comparative experiment. We can see that, unlike REGIS, on CHAVE, both correction systems improved results in all metrics. We attribute this difference to two reasons (i) the size of the documents: it seems well established in the literature that shorter documents may benefit more from correction (CROFT et al., 1994; TAGHVA et al., 1994) ; (ii) the error rates in the baseline run in CHAVE were significantly higher – *i.e.,* a 25% WER compared to a 6.49% WER in REGIS, which mean a larger room for improvements.

Table 6.6 – Error correction results obtained in REGIS and CHAVE collections with 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%

Source: The Authors

7 CONCLUSION

In this work, we presented REGIS, an IR test collection for the geoscientific domain in Portuguese. We described the entire process followed for document collection, topic creation, and the generation of relevance assessments. Using REGIS, we also analyzed the impacts of OCR extraction and error correction in information retrieval, using three tools (ABBYY, Tika, and Tornado) to extract the textual contents of the documents and two correction methods (`sOCRates` and SymSpell) to post-process the text.

In a language that, to this date, has only a single test collection, we believe REGIS can help foment IR research. It can be used to assess a variety of techniques, including solutions for automatic query reformulation, stemming, query expansion, and scoring functions. Since the original documents are in PDF, we used REGIS to test the impact that correcting OCR extraction errors has on IR results.

Our experiments with three OCR tools (Section 6.1) 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, although it has the disadvantage of being a proprietary software that also poses a limit on the number of pages that can be extracted by month. The documents in REGIS are very long (thesis, dissertations, and technical reports, averaging 25K tokens) and yet they were significantly impacted by OCR errors. This finding contradicts existing work that suggested that long documents were robust to these errors (CROFT et al., 1994; TAGHVA et al., 1994; MITTENDORF; SCHÄUBLE, 2000).

Existing work on the impact of OCR error correction on IR metrics have reached contrasting conclusions. While some works pointed that post-processing can improve IR quality (EVERSHED; FITCH, 2014; VARGAS et al., 2021; ZHUANG; ZUCCON, 2021) others found the opposite (TAGHVA et al., 1994; CROFT et al., 1994; MITTENDORF; SCHÄUBLE, 2000). 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 (Section 6.4). That required the manual creation of a ground truth for the text extraction process. The results also confirm that ABBYY was the best text extraction tool with an estimated word error rate of 1.57. The second-best performer, Tika, had 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. According to the figures reported by (HEGGHAMMER, 2021), it would cost between 3,600 and 145,000 US dollars to process the entire REGIS collection with these cloud tools[1]. 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.

Finally, limitation in our intrinsic evaluation is the small number of samples in our ground truth dataset. Although our ground truth is not a representative sample for REGIS collection, the experiments with this sample allowed important insights over the text quality from each method. Extra annotation 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 the context of this MSc. dissertation, we published a paper describing REGIS at SIGIR 2021 (OLIVEIRA; ROMEU; MOREIRA, 2021) and submitted an article to Information Processing & Management describing the evaluation of OCR extraction and correction and its impact on IR. Additionally, I was co-author of the paper that proposes `sOCRates` (VARGAS et al., 2021).

As future work, there are many possible directions to follow. One of them is to generate finer-grained annotation for REGIS, where instead of the complete document, the most relevant excerpts in the document that answer each query would be identified. Also, one could assess the impact of other methods, such as relevance feedback, query expansion and test new correction methods as the TrOCR (LI et al., 2021), recently released by Microsoft researchers. Dedicate extra effort to generate an ideal test collection, as described in Section 5.1, could also be pursued. This resource could be achieved by generating a representative ground truth from REGIS.

[1] REGIS documents have a total of 2.4 million pages. The costs mentioned by (HEGGHAMMER, 2021) range between $1.5 and 60 US dollars per 1000 pages

REFERENCES

BASU, M. et al. Microblog retrieval in a disaster situation: A new test collection for evaluation. In: **SMERP@ ECIR**. [S.l.: s.n.], 2017. p. 22–31.

BAZZO, G. T. et al. Assessing the impact of OCR errors in information retrieval. In: **European Conference on Information Retrieval**. [S.l.: s.n.], 2020. p. 102–109.

BENDER, E. M. **English isn't generic for language, despite what NLP papers might lead you to believe**. 2019. Available from Internet: <https://faculty.washington.edu/ebender/papers/Bender-SDSS-2019.pdf>.

BUCKLEY, C.; VOORHEES, E. M. Retrieval evaluation with incomplete information. In: **Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval**. [S.l.: s.n.], 2004. p. 25–32.

BUCKLEY, C.; VOORHEES, E. M. Evaluating evaluation measure stability. In: **ACM SIGIR Forum**. [S.l.: s.n.], 2017. v. 51, p. 235–242.

CARRASCO, R. C. An open-source OCR evaluation tool. In: **Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage**. [S.l.: s.n.], 2014. p. 179–184.

CHIRON, G. et al. ICDAR2017 competition on post-OCR text correction. In: **2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)**. [S.l.: s.n.], 2017. v. 1, p. 1423–1428.

CLEVERDON, C. W. **Aslib Cranfield research project: report on the testing and analysis of an investigation into the comparative efficiency of indexing systems**. [S.l.], 1962.

CONSOLI, B. et al. Embeddings for named entity recognition in geoscience portuguese literature. In: **Proceedings of The 12th Language Resources and Evaluation Conference**. [S.l.: s.n.], 2020. p. 4625–4630.

CORDEIRO, F. C.; VILLALOBOS, C. E. M. Petrolês - how to build a specialized oil and gas corpus in portuguese. **Rio Oil and Gas Expo and Conference**, v. 20, n. 2020, p. 387–388, dec. 2020.

CROFT, W. B. et al. An evaluation of information retrieval accuracy with simulated OCR output. In: **Symposium on Document Analysis and Information Retrieval**. [S.l.: s.n.], 1994. p. 115–126.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

DROBAC, S.; LINDÉN, K. Optical character recognition with neural networks and post-correction with finite state methods. **International Journal on Document Analysis and Recognition (IJDAR)**, v. 23, n. 4, p. 279–295, 2020.

DUTTA, H.; GUPTA, A. PNRank: Unsupervised ranking of person name entities from noisy OCR text. **Decision Support Systems**, Elsevier, v. 152, p. 113662, 2022.

EVERSHED, J.; FITCH, K. Correcting noisy OCR: Context beats confusion. In: **Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage**. [S.l.: s.n.], 2014. p. 45–51.

FLORES, F. N.; MOREIRA, V. P. Assessing the impact of stemming accuracy on information retrieval–a multilingual perspective. **Information Processing & Management**, v. 52, n. 5, p. 840–854, 2016.

GHOSH, K. et al. Improving information retrieval performance on OCRed text in the absence of clean text ground truth. **Information Processing & Management**, v. 52, n. 5, p. 873–884, 2016.

GOMES, D. et al. Portuguese word embeddings for the oil and gas industry: Development and evaluation. **Computers in Industry**, v. 124, p. 103347, 2021. ISSN 0166-3615.

GUPTE, A. et al. **Lights, Camera, Action! A Framework to Improve NLP Accuracy over OCR documents**. 2021.

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)**. [S.l.: s.n.], 2019. p. 431–436.

HAMDI, A. et al. Assessing and minimizing the impact of OCR quality on named entity recognition. In: SPRINGER. **International Conference on Theory and Practice of Digital Libraries**. [S.l.], 2020. p. 87–101.

HEGGHAMMER, T. OCR with tesseract, amazon textract, and google document ai: a benchmarking experiment. **Journal of Computational Social Science**, Springer, p. 1–22, 2021.

HERSH, W. et al. Do batch and user evaluations give the same results? In: **Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval**. [S.l.: s.n.], 2000. p. 17–24.

HUYNH, V.-N.; HAMDI, A.; DOUCET, A. When to use OCR post-correction for named entity recognition? In: SPRINGER. **International Conference on Asian Digital Libraries**. [S.l.], 2020. p. 33–42.

JIANG, M. et al. Impact of OCR quality on bert embeddings in the domain classification of book excerpts. **Proceedings http://ceur-ws. org ISSN**, v. 1613, p. 0073, 2021.

JING, H.; LOPRESTI, D.; SHIH, C. Summarization of noisy documents: a pilot study. In: **Proceedings of the HLT-NAACL 03 Text Summarization Workshop**. [S.l.: s.n.], 2003. p. 25–32.

JOHNSON, S. et al. Spoken document retrieval for TREC-7 at cambridge university. In: **TREC**. [S.l.: s.n.], 1999. v. 7, p. 1.

KANTOR, P. B.; VOORHEES, E. M. The TREC-5 confusion track: Comparing retrieval methods for scanned text. **Information Retrieval**, v. 2, n. 2, p. 165–176, May 2000. ISSN 1573-7659.

LAM-ADESINA, A. M.; JONES, G. J. Examining and improving the effectiveness of relevance feedback for retrieval of scanned text documents. **Information Processing & Management**, v. 42, n. 3, p. 633–649, 2006.

LI, M. et al. TrOCR: Transformer-based optical character recognition with pre-trained models. **arXiv preprint arXiv:2109.10282**, 2021.

LIN, X. Impact of imperfect OCR on part-of-speech tagging. In: **Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.** [S.l.: s.n.], 2003. p. 284–288.

LYKKE, M. et al. Developing a test collection for the evaluation of integrated search. In: **European Conference on Information Retrieval**. [S.l.: s.n.], 2010. p. 627–630.

MANNING, C. D.; RAGHAVAN, P.; SCHüTZE, H. Evaluation in information retrieval. In: ____. **Introduction to Information Retrieval**. [S.l.: s.n.], 2008. p. 139–161.

MEI, J. et al. Statistical learning for OCR error correction. **Information Processing & Management**, v. 54, n. 6, p. 874–887, 2018.

MEMON, J. et al. Handwritten optical character recognition (ocr): A comprehensive systematic literature review (slr). **IEEE Access**, IEEE, v. 8, p. 142642–142668, 2020.

MILLER, D. et al. Named entity extraction from noisy input: speech and OCR. In: **Sixth Applied Natural Language Processing Conference**. [S.l.: s.n.], 2000. p. 316–324.

MITTENDORF, E.; SCHÄUBLE, P. Information retrieval can cope with many errors. **Information Retrieval**, v. 3, n. 3, p. 189–216, 2000.

MUTUVI, S. et al. Evaluating the impact of OCR errors on topic modeling. In: **International Conference on Asian Digital Libraries**. [S.l.: s.n.], 2018. p. 3–14.

NGUYEN, T. et al. Deep statistical analysis of OCR errors for effective post-OCR processing. In: **Joint Conference on Digital Libraries (JCDL)**. [S.l.: s.n.], 2019. p. 29–38.

NGUYEN, T. T. H. et al. Survey of post-OCR processing approaches. **ACM Computing Surveys (CSUR)**, v. 54, n. 6, p. 1–37, 2021.

OLIVEIRA, L. Lima de; 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**. [S.l.: s.n.], 2021. p. 2363–2368. ISBN 9781450380379.

PONTES, E. L. et al. Impact of OCR quality on named entity linking. In: SPRINGER. **International Conference on Asian Digital Libraries**. [S.l.], 2019. p. 102–115.

RIGAUD, C. et al. ICDAR 2019 competition on post-OCR text correction. In: **2019 International Conference on Document Analysis and Recognition (ICDAR)**. [S.l.: s.n.], 2019. p. 1588–1593.

RITCHIE, A.; TEUFEL, S.; ROBERTSON, S. Creating a test collection for citation-based IR experiments. In: **Proceedings of the human language technology conference of the NAACL, main conference**. [S.l.: s.n.], 2006. p. 391–398.

SANDERSON, M. **Test collection based evaluation of information retrieval systems**. [S.l.: s.n.], 2010.

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**. [S.l.: s.n.], 2004. p. 821–832.

SINGH, S. Optical character recognition techniques: a survey. **Journal of emerging Trends in Computing and information Sciences**, v. 4, n. 6, p. 545–550, 2013.

SOBOROFF, I.; GRIFFITT, K.; STRASSEL, S. The BOLT IR test collections of multilingual passage retrieval from discussion forums. In: **Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval**. [S.l.: s.n.], 2016. p. 713–716.

SPARK-JONES, K. Report on the need for and provision of an 'ideal' information retrieval test collection. **Computer Laboratory**, 1975.

STRIEN, D. van 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**. [S.l.: s.n.], 2020. p. 484–496.

TAGHVA, K.; BORSACK, J.; CONDIT, A. Effects of OCR errors on ranking and feedback using the vector space model. **Information Processing & Management**, v. 32, n. 3, p. 317–327, 1996.

TAGHVA, K.; BORSACK, J.; CONDIT, A. Evaluation of model-based retrieval effectiveness with OCR text. **ACM Transactions on Information Systems (TOIS)**, v. 14, n. 1, p. 64–93, 1996.

TAGHVA, K. et al. The effects of noisy data on text retrieval. **Journal of the American Society for Information Science**, v. 45, n. 1, p. 50–58, 1994.

TRAUB, M. C. et al. Impact of crowdsourcing OCR improvements on retrievability bias. In: **Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries**. [S.l.: s.n.], 2018. p. 29–36.

TURPIN, A.; SCHOLER, F. User performance versus precision measures for simple search tasks. In: **Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval**. [S.l.: s.n.], 2006. p. 11–18.

VARGAS, D. S. et al. sOCRates-a post-OCR text correction method. In: **Anais do XXXVI Simpósio Brasileiro de Bancos de Dados**. [S.l.: s.n.], 2021. p. 61–72.

WIEDENHOFER, L.; HEIN, H.-G.; DENGEL, A. Post-processing of OCR results for automatic indexing. In: IEEE. **Proceedings of 3rd International Conference on Document Analysis and Recognition**. [S.l.], 1995. v. 2, p. 592–596.

YILMAZ, E.; ASLAM, J. A. Estimating average precision with incomplete and imperfect judgments. In: **Proceedings of the 15th ACM international conference on Information and knowledge management**. [S.l.: s.n.], 2006. p. 102–111.

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**. [S.l.: s.n.], 2021. p. 2836–2842.

ZU, G. et al. The impact of OCR accuracy on automatic text classification. In: **Advanced Workshop on Content Computing**. [S.l.: s.n.], 2004. p. 403–409.

APPENDIX A — DOCUMENTS FROM REGIS AND CHAVE COLLECTIONS

Figure A.1 – Examples of documents from the test collection

<doc>
<field name="docid"> BR-BG.03782 < \field>
<field name="filename"> BGP_1990_4_4_05_Preservacao_[...].pdf < \field>
<field name="filetype"> PDF < \field>
<field name="text"> [...]
RESUMO - A preservação e a geração de porosidade em reservatórios elásticos profundos são controladas por diversos processos e situações geológicas específicas. Os principais fatores de preservação de porosidade são os seguintes: 1)- soterramento tardio do reservatório à sua atual profundidade; 2)- desenvolvimento de pressões anormais de fluidos; 3)- estabilidade composicional dos grãos do arcabouço; 4)-recobrimento dos grãos por cutículas ou franjas de argilas e/ou óxidos; 5)- cimentação precoce parcial por carbonatos ou sulfatos; e 6)- saturação precoce do reservatório por hidrocarbonetos. Os processos e solventes para a geração de porosidade em subsuperffcie são estes: 1)- infiltração profunda de águas meteóricas; 2)- CO2 da maturação térmica da matéria orgânica; 3)- solventes orgânicos (principalmente ácidos carboxflicos) liberados pela matéria orgânica; 4)- fluidos ácidos de reações inorgânicas com argilominerais; 5)- redução termogênica de sulfato por hidrocarbonetos, produzindo CO2 e H2S; 6)- convecção térmica de fluidos solventes; 7)- superposição de geradores associados ao mesmo reservatório; 8)- mistura de águas meteóricas com águas marinhas ou conatas; 9|- complexos inorgânicos com cloro; 10)- amônia; c11)-águas "juvenis" com CO2 de fontes hidrotermais, vulcânicas, ou do metamorfismo de calcários. Um balanço dos mecanismos de preservação indica que a saturação precoce do reservatório por hidrocarbonetos seja a mais eficiente, embora o soterramento tardio seja provavelmente o de mais ampla influência. Entre os processos de geração de porosidade, os solventes orgânicos ainda parecem ser os mais importantes na geração de porosidade em subsuperffcie, mas diversos outros processos podem ser muito influentes, devendo ser também sistematicamente avaliados. [...] < \field>
< \doc>

(a) Excerpt of a Document from REGIS

<doc>
<field name="docno"> FSP950101-036 < \field>
<field name="docid"> FSP950101-036 < \field>
<field name="date"> 950101 < \field>
<field name="category"> MUNDO < \field>
<field name="text"> Da Reportagem Local Passado o período de festas, a dona-de-casa deve ficar atenta para mais uma despesa. É que o governo decidiu na semana passada pagar um abono de R$ 15,00 para quem ganha salário mínimo. Com a medida, que é válida a partir deste mês, nenhum trabalhador no país pode receber salário inferior a R$ 85,00 (US$ 100). O empregado doméstico, que tem direito por lei a ganhar pelo menos um salário mínimo, também terá que receber esse abono. O valor de R$ 15,00 será pago apenas no mês de janeiro e não será incorporado ao salário mensal. O pagamento deve ser feito até o quinto dia útil de fevereiro (dia 6). Mas atenção. Só tem direito ao abono quem ganha um salário mínimo em dezembro. Suponha que você já pague mais de R$ 85,00 para sua empregada. Nesse caso, não terá que dar qualquer abono. Quem registrou na carteira profissional do doméstico seu salário em quantidade de salários mínimos também deve pagar o abono. < \field>
< \doc>

(b) Document from CHAVE

APPENDIX B — RESUMO EXPANDIDO EM PORTUGUÊS: CRIANDO RECURSOS E AVALIANDO O IMPACTO DA QUALIDADE DO OCR NA RECUPERAÇÃO DA INFORMAÇÃO: UM ESTUDO DE CASO NO DOMÍNIO GEOCIENTÍFICO

O Formato de Documento Portátil (PDF) se tornou um dos padrões mais usados para armazenamento e compartilhamento de documentos. Artigos científicos, propostas de projetos, contratos, livros e documentos jurídicos são normalmente armazenados e distribuídos como arquivos PDF. Embora a extração do conteúdo textual de documentos PDF originados de forma digital possa ser feita com alta precisão, se o documento consistir em uma imagem digitalizada, o Reconhecimento Óptico de Caracteres (OCR) é normalmente necessário. A saída do OCR pode ser ruidosa, especialmente quando a qualidade da imagem digitalizada é ruim – muito comum em documentos históricos –, o que por sua vez pode impactar tarefas posteriores, como Recuperação de Informação (IR).

O pós-processamento de documentos OCR é uma alternativa para corrigir erros de extração e, intuitivamente, melhorar os resultados em tarefas posteriores. Este trabalho avalia o impacto da extração e correção de OCR em IR. Comparamos diferentes métodos de extração e correção em textos extraídos por OCR de documentos escaneados reais.

Para avaliar as tarefas de IR, o paradigma padrão requer uma coleção de testes com documentos, consultas e julgamentos de relevância. A criação de coleções de teste requer um esforço humano significativo, principalmente na realização dos julgamentos de relevância. Como resultado, ainda existem muitos domínios e idiomas que, até hoje, carecem de um ambiente de teste para avaliação adequada. O português é um exemplo de uma importante língua mundial que tem sido negligenciada em termos de pesquisas de IR - a única coleção de testes disponível é composta por notícias de 1994 e uma centena de consultas.

Com o objetivo de preencher essa lacuna, desenvolvemos a REGIS (*Retrieval Evaluation for Geoscientific Information Systems*), uma coleção de testes para o domínio geocientífico em português. REGIS contém 20 mil documentos e 34 tópicos de consulta, juntamente com julgamentos de relevância. Neste trabalho descrevemos os procedimentos para a coleta de documentos, criação de tópicos e geração dos julgamentos de relevância. Para facilitar nesse processo e dar suporte durante toda a criação da colução, foi desenvolvimento um sistema de anotação. Contamos com a ajuda de especialistas

do domínios de geociências para fornecer os julgamentos sobre os pares de consultas e documentos.

Com a coleção REGIS à disposição, realizamos uma série de experimentos para avaliar o desempenho das diferentes versões de texto disponíveis. Ao todo foram cinco versões, sendo três delas de extratores OCR (Apache Tika, ABBYY FineReader e Tornado) e duas delas de métodos de pós processamento (SymSpell e `sOCRates`). Os experimentos dos nossos resultados mostraram que, em média, para o conjunto completo de tópicos de consulta, as métricas de qualidade de recuperação variam muito pouco. No entanto, uma análise mais detalhada revelou que a maioria dos tópicos de consulta melhorou com a correção de erros.

Dentre as contribuições deste trabalho, podemos citar: (i) a criação de um sistema de anotações completo para auxiliar na creação de coleções de teste e obtenção de julgamentos de relevancia, (ii) criação de uma nova coleção de testes de IR em português, (iii) uma investigação do impacto de diferentes métodos de extração e correção para textos obtidos via OCR usando documentos reais, (iv) uma avaliação da qualidade intrínseca da extração de texto e correção de erros, (v) realização de experimentos com um idioma que, apesar de amplamente falado, é pouco representado em termos de recursos de IR.

Como trabalhos futuros, há muitas direções possíveis a serem seguidas. Uma delas é gerar anotação mais granualar para a REGIS, onde ao invés do documento completo, seriam identificados os trechos mais relevantes do documento que respondem a cada consulta. Também pretendemos avaliar o impacto de outros métodos de extração e correção.