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

Lucas Lima de Oliveira, Viviane P. Moreira

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{lloliveira,viviane}@inf.ufrgs.br

Abstract. *The evaluation paradigm in Information Retrieval (IR) requires a test collection with documents, queries, and relevance judgments. Creating such collections demands significant human effort, mainly to provide relevance judgments. As a result, there are still many domains and languages that, to this day, lack a proper evaluation testbed. To bridge this gap, we developed REGIS (Retrieval Evaluation for Geoscientific Information Systems), a test collection for the geoscientific domain in Portuguese. The documents in REGIS are in PDF. Optical Character Recognition (OCR) is typically used to extract the textual contents of scanned texts. The output of OCR can be noisy, especially when the quality of the scanned image is poor, which in turn can impact downstream tasks such as Information Retrieval. This work evaluates the impact of OCR extraction and correction on IR. Our results have shown significant differences in IR metrics for the different digitization methods.*

Resumo. *The evaluation paradigm in Information Retrieval (IR) requires a test collection with documents, queries, and relevance judgments. Creating such collections demands significant human effort, mainly to provide relevance judgments. As a result, there are still many domains and languages that, to this day, lack a proper evaluation testbed. To bridge this gap, we developed REGIS (Retrieval Evaluation for Geoscientific Information Systems), a test collection for the geoscientific domain in Portuguese. The documents in REGIS are in PDF. Optical Character Recognition (OCR) is typically used to extract the textual contents of scanned texts. The output of OCR can be noisy, especially when the quality of the scanned image is poor, which in turn can impact downstream tasks such as Information Retrieval. This work evaluates the impact of OCR extraction and correction on IR. Our results have shown significant differences in IR metrics for the different digitization methods.*

1. Data about the MSc Thesis

- **Defense date:** 27/01/2022
- **Program/University:** Programa de Pós Graduação em Computação (PPGC)
  Category: Master's degree
- **Evaluation panel:** Carina Friedrich Dorneles (UFSC), Mara Abel (UFRGS), Leandro Krug Wives (UFRGS)

2. Context

A significant part of textual information exchanged in digital documents is typically stored and distributed in Portable Document Format (PDF). Before being fed to Information Retrieval (IR) algorithms, the textual contents of PDF files need to be extracted. When the file was not digitally created, 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. The impact that OCR errors have on retrieval quality is still an open question.

To test the quality of any IR solution, a test collection consisting of documents, queries, and relevance judgments is necessary. Given their importance, significant effort has been devoted to building test collections since the early days of IR research. 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 that has been overlooked in terms of linguistic resources. The only existing standard 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 [Santos and Rocha 2004].

The Oil and Gas (O&G) industry plays an important role in Portuguese-speaking countries, representing an essential part of their economies. Despite the importance of this industry, there are few linguistic resources available for this sub-domain of the Geosciences.

3. Goals

The aim of this work was two-fold. First, it intends to address the lack of IR collections in Portuguese, more specifically, in the Geoscientific domain. Then, using the test collection, we expect to answer two main research questions: (i) How does the quality of the digitization affect retrieval results? and (ii) Can post-processing OCR-ed texts improve retrieval quality?

4. Contributions

Our first contribution is the REGIS collection[1] (Retrieval Evaluation for Geoscientific Information Systems); it is composed of over 20 thousand documents, 34 query topics, and their corresponding relevance judgments using a four-level scale ("very relevant", "fairly relevant", "marginally relevant", and "not relevant"). The documents were produced over a long time span (1957 to 2020) and vary substantially in terms of visual quality. As a byproduct, we developed an annotation system that allows the complete set of CRUD (create, read, update, and delete) operations over queries, documents, and relevance judgments. The code of the annotation system was released[2] and may be employed by other researchers in similar annotation tasks. Our second contribution is an investigation of the impact of different text extraction and correction methods for OCR-ed texts using real OCR-ed data.

[1] REGIS: `https://github.com/Petroles/regis-collection`

[2] `https://github.com/lucaslioli/regis-system`

5. Advancement of the state-of-the-art

Although the impacts of OCR-ed text in IR have already been the focus of a few studies [Croft et al. 1994, Taghva et al. 1996b, Kantor and Voorhees 2000, Ghosh et al. 2016, Bazzo et al. 2020], there is not much work on evaluating the impact of post-processing techniques that try to fix digitization errors. Previous work [Vargas 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. Contrary to existing work that argues that long documents are robust to OCR errors [Croft et al. 1994, Taghva et al. 1996a, Mittendorf and Schäuble 2000], we found that retrieval quality metrics varied significantly depending on the digitization system. For error correction, our results showed that on average for the complete set of query topics, retrieval quality metrics change very little. However, a more detailed analysis showed that most query topics (19 out of 34) improved with error correction.

In a language that, up to this date, had only a single standard test collection, 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.

The creation of REGIS followed the best practices in test collection development. We adopted the *pooling method* proposed by [Spark-Jones 1975] and well-described by [Sanderson 2010], which is the standard for this type of resource. We relied on the collaboration of domain specialists to create query topics that mimic real user needs and judge the relevance of documents with respect to the queries. REGIS has a broad range of topics to ensure a mix between generic and specific queries, as well as easier and harder ones.

Using REGIS, we analyzed the impacts of OCR errors and error correction in IR. We start by taking the original PDF documents and extracting their textual contents. Once the text is extracted, it is submitted to a post-processing step that aims to fix digitization errors. Then, at the evaluation stage, the queries are run and scored using the relevance judgments. We also carried out an *intrinsic* evaluation to provide insights into the inherent quality of the digitization and correction processes, regardless of the retrieval results. Such an evaluation required the creation of ground truth digitizations.

6. Evaluation

Our experiments compared three digitization tools (Apache Tika[3], ABBYY FineReader[4], and Tornado[5]) and two error correction methods (sOCRates [Vargas et al. 2021] and SymSpell[6]). To measure their impact on retrieval quality, we computed *mean average precision* (MAP) and Normalized Discounted Cumulative Gain (*NDCG*), which are the standard metrics to evaluate ranked results.

We found statistically significant differences in retrieval quality. ABBYY, the best performer, yielded MAP scores that were about 4 percentage points higher than Tika's,

[3] https://tika.apache.org/
[4] https://www.abbyy.com/
[5] https://petroles.puc-rio.ai/index_en.html
[6] https://github.com/wolfgarbe/SymSpell

but it has the disadvantage of being a proprietary software that also poses a limit on the number of pages that can be extracted.

The results of the intrinsic evaluation also confirm that ABBYY was the best digitization tool with an estimated word error rate of 1.62, while Tika and Tornado obtained an error rate four times greater. Still, budget constraints may deter the adoption of paid tools such as ABBYY, Amazon Textract, or Google Document AI. 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. 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.

7. Scientific and Technical Production

Two papers were published as direct results of this M.Sc. thesis. Both venues are rated as **Qualis A1** by CAPES.

- [Oliveira et al. 2021][7] – a resource paper at ACM SIGIR, which is the most prominent conference in IR, describing the development of the REGIS test collection.
- [Oliveira et al. 2023][8] – an article published in the International Journal on Digital Libraries describing the investigation of the impact of OCR digitization and correction in IR.

Four technical products were produced as direct contributions of this work.

- REGIS collection[9] – acronym for Retrieval Evaluation for Geoscientific Information Systems, it is an IR test collection for the geoscientific domain in Portuguese. REGIS contains 20K documents and 34 query topics, along with relevance assessments.
- REGIS Annotation system[10] – is a system to enable the annotation process and generate relevance judgments for documents over queries. The system supports the entire CRUD (create, read, update, and delete) operations over queries, documents, and annotations. It also includes dashboards to monitor the progress of annotations.
- Ground truth annotations for OCR digitization[11] – repository containing a sample of 170 sentences, with an average of 30 tokens by sentence and their corresponding pages from the REGIS collection manually collected and corrected by two annotators. This small dataset can be used to assess the quality of OCR systems.
- Code for Apache Solr experiments[12] – script to submit a set of queries to an indexed collection in Apache Solr, based on an XML file. The output is a ranking containing the top 100 most relevant documents to each topic. Each query is searched using proximity search, and it is also possible to determine the usage of description and narrative to search.

Acknowledgments. This work has been partially funded by CENPES Petrobras, CNPq-Brazil, and Capes Finance Code 001.

[7] https://doi.org/10.1145/3404835.3463256
[8] https://doi.org/10.1007/s00799-023-00345-6
[9] https://zenodo.org/records/4726013
[10] https://github.com/lucaslioli/regis-system
[11] https://github.com/lucaslioli/regis-collection-gs
[12] https://github.com/lucaslioli/solr-query-script

References

Bazzo, G. T., Lorentz, G. A., Vargas, D. S., and Moreira, V. P. (2020). Assessing the impact of OCR errors in information retrieval. In *European Conference on Information Retrieval*, pages 102–109.

Croft, W. B., Harding, S., Taghva, K., and Borsack, J. (1994). An evaluation of information retrieval accuracy with simulated OCR output. In *Symposium on Document Analysis and Information Retrieval*, pages 115–126.

Ghosh, K., Chakraborty, A., Parui, S. K., and Majumder, P. (2016). Improving information retrieval performance on OCRed text in the absence of clean text ground truth. *Information Processing & Management*, 52(5):873–884.

Hegghammer, T. (2021). OCR with tesseract, amazon textract, and google document ai: a benchmarking experiment. *Journal of Computational Social Science*, pages 1–22.

Kantor, P. B. and Voorhees, E. M. (2000). The TREC-5 confusion track: Comparing retrieval methods for scanned text. *Information Retrieval*, 2(2):165–176.

Mittendorf, E. and Schäuble, P. (2000). Information retrieval can cope with many errors. *Information Retrieval*, 3(3):189–216.

Oliveira, L. L. d., Romeu, R. K., and Moreira, V. P. (2021). 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*, page 2363–2368.

Oliveira, L. L. d., Vargas, D. S., Alexandre, A. M. A., Cordeiro, F. C., Gomes, D. d. S. M., Rodrigues, M. d. C., Romeu, R. K., and Moreira, V. P. (2023). Evaluating and mitigating the impact of OCR errors on information retrieval. *International Journal on Digital Libraries*, 24(1):45–62.

Sanderson, M. (2010). Test collection based evaluation of information retrieval systems. *Foundations and Trends® in Information Retrieval*, 4(4):247–375.

Santos, D. and Rocha, P. (2004). 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*, pages 821–832.

Spark-Jones, K. (1975). Report on the need for and provision of an 'ideal' information retrieval test collection. *Computer Laboratory*.

Taghva, K., Borsack, J., and Condit, A. (1996a). Effects of OCR errors on ranking and feedback using the vector space model. *Information Processing & Management*, 32(3):317–327.

Taghva, K., Borsack, J., and Condit, A. (1996b). Evaluation of model-based retrieval effectiveness with OCR text. *ACM Transactions on Information Systems (TOIS)*, 14(1):64–93.

Vargas, D. S., de Oliveira, L. L., Moreira, V. P., Bazzo, G. T., and Lorentz, G. A. (2021). sOCRates-a post-OCR text correction method. In *Anais do XXXVI Simpósio Brasileiro de Bancos de Dados*, pages 61–72.