REGIS: A Test Collection for Geoscientific Documents in Portuguese

Lucas Lima de Oliveira
Institute of Informatics, UFRGS
Porto Alegre, Brazil
lloliveira@inf.ufrgs.br

Regis Kruel Romeu
CENPES/Petrobras
Rio de Janeiro, Brazil
regisromeu@gmail.com

Viviane Pereira Moreira
Institute of Informatics, UFRGS
Porto Alegre, Brazil
viviane@inf.ufrgs.br

ABSTRACT

Experimental validation is key to the development of Information Retrieval (IR) systems. The standard evaluation 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, in this paper, 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. We describe the procedures for document collection, topic creation, and relevance assessment. In addition, we report on results of standard IR techniques on REGIS so that they can serve as a baseline for future research.

CCS CONCEPTS

• **Information systems → Test collections**.

KEYWORDS

Test collection, information retrieval, geoscientific data

ACM Reference Format:
Lucas Lima de Oliveira, Regis Kruel Romeu, and Viviane Pereira Moreira. 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 (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3404835.3463256

1 INTRODUCTION

The digitization of industries brought many new challenges to information retrieval (IR), 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 [5]. These efforts were intensified in the 90s in the US, with the Text REtrieval Conferences (TREC)[1]. Similar efforts in Europe (CLEF)[2] and Asia (NTCIR)[3] 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 [14].

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.* [8], despite the importance of this industry, 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 [8]. These documents incorporate a highly technical vocabulary, which is not covered by the available IR test collections, especially in Portuguese. This limitation poses difficulties for Brazilian researchers to test their methods and for companies to implement them internally.

In an attempt to address this gap, in this paper, we propose a test collection for the geoscientific domain in Portuguese. The collection is called REGIS (Retrieval Evaluation for Geoscientific Information Systems); it is composed of over 20 thousand documents, 34 query topics, and their corresponding relevance judgments. REGIS was created with the cooperation of domain specialists, following the pooling method proposed by Spärck-Jones & Van Rijsbergen [16] and well described by Sanderson [13]. REGIS collection is available at https://github.com/Petroles/regis-collection.

The remainder of this paper is organized as follows. Section 2 presents related work on similar datasets. Section 3.1 presents

SIGIR '21, July 11–15, 2021, Virtual Event, Canada.
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3463256

[1] TREC: https://trec.nist.gov/
[2] CLEF: http://www.clef-initiative.eu/
[3] NTCIR: http://ntcir.nii.ac.jp/

REGIS, detailing the processes adopted in data collection, topic building, pool creation, and how the relevance judgments were obtained. Section 4 reports on the results of standard retrieval methods to serve as baselines for future work. Finally, Section 6 concludes this paper.

2 RELATED WORK

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 [14], 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.* [10] 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 *et al.* [12] created a test collection with scientific papers from the ACL anthology. The goal was to explore evidences 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.* [1] 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 *et al.* [15] 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 [7], word embeddings [8], and named-entity recognition systems [6]. 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 THE REGIS IR TEST 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 recommendations by Manning *et al.* [11] and Sanderson [13]. 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.

Table 1: REGIS Statistics

Theses and Dissertations	19,489
Bulletins Technical reports	1,955
Total Documents	21,444
Tokens	538.4M
Distinct tokens	4.1M
Avg Tokens per doc	25.1K

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.

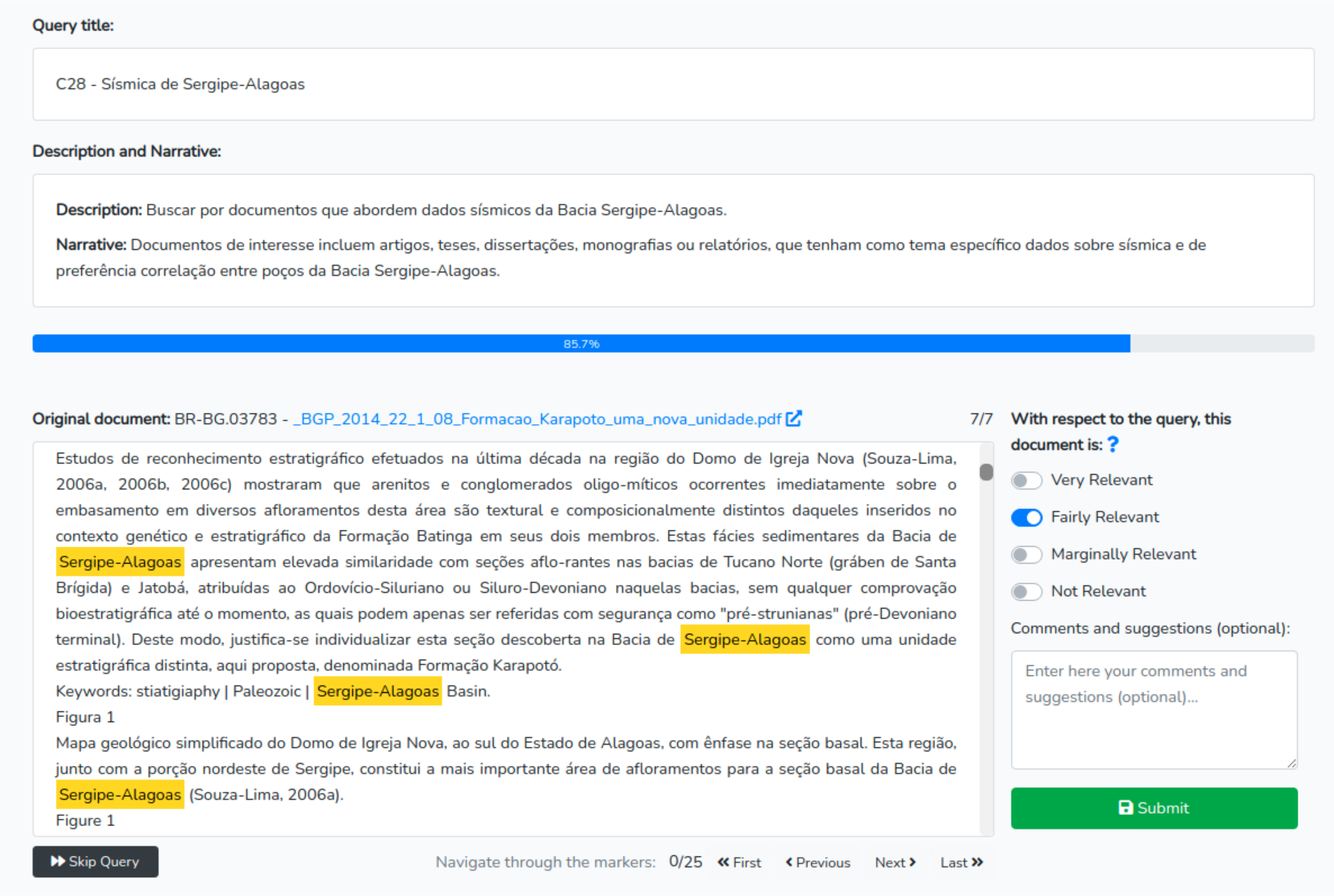
- `docid`: Document unique identification.
- `filename`: Name of the original file.
- `filetype`: Type of the original file.
- `text`: Document contents, extracted from the original PDF file.

3.1 Data collection and 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)[4] and a Brazilian Oil Company[5]. 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 Institute of Information in Science and Technology (IBICT)[6]. 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 (Optical Character Recognition), namely ABBYY FineReader[7], and Tika[8]. Then, duplicate documents were detected and removed. Finally, a total of 21,444 documents remained. Details of the documents are shown in Table 1. 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] ANP: https://www.gov.br/anp/pt-br/
[5] Petrobras: http://publicacoes.petrobras.com.br/
[6] IBICT: https://bdtd.ibict.br/vufind/
[7] ABBYY: https://www.abbyy.com/
[8] Tika: https://tika.apache.org/



Query title:

C28 - Sísmica de Sergipe-Alagoas

Description and Narrative:

Description: Buscar por documentos que abordem dados sísmicos da Bacia Sergipe-Alagoas.

Narrative: 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.

85.7%

Original document: BR-BG.03783 - _BGP_2014_22_1_08_Formacao_Karapoto_uma_nova_unidade.pdf 7/7

Estudos de reconhecimento estratigráfico efetuados na última década na região do Domo de Igreja Nova (Souza-Lima, 2006a, 2006b, 2006c) mostraram que arenitos e conglomerados oligo-míticos ocorrentes imediatamente sobre o embasamento em diversos afloramentos desta área são textural e composicionalmente distintos daqueles inseridos no contexto genético e estratigráfico da Formação Batinga em seus dois membros. Estas fácies sedimentares da Bacia de Sergipe-Alagoas apresentam elevada similaridade com seções aflo-rantes nas bacias de Tucano Norte (gráben de Santa Brígida) e Jatobá, atribuídas ao Ordovício-Siluriano ou Siluro-Devoniano naquelas bacias, sem qualquer comprovação bioestratigráfica até o momento, as quais podem apenas ser referidas com segurança como "pré-strunianas" (pré-Devoniano terminal). Deste modo, justifica-se individualizar esta seção descoberta na Bacia de Sergipe-Alagoas como uma unidade estratigráfica distinta, aqui proposta, denominada Formação Karapotó.
Keywords: stiatigiaphy | Paleozoic | Sergipe-Alagoas Basin.
Figura 1
Mapa geológico simplificado do Domo de Igreja Nova, ao sul do Estado de Alagoas, com ênfase na seção basal. Esta região, junto com a porção nordeste de Sergipe, constitui a mais importante área de afloramentos para a seção basal da Bacia de Sergipe-Alagoas (Souza-Lima, 2006a).
Figura 1

With respect to the query, this document is: ?

○ Very Relevant
● Fairly Relevant
○ Marginally Relevant
○ Not Relevant

Comments and suggestions (optional):

Enter here your comments and suggestions (optional)...

Figure 1: Screenshot of the annotation page of the system

3.2 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.
- `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 2 shows an example of a query topic.

3.3 Pool Creation

Building a test collection in which annotators judge the query-document pairs is unfeasible. To address this problem, the pooling methodology was proposed by Spärck-Jones & Van Rijsbergen [16] and became the standard procedure. The principle of this methodology is to create a smaller subset with representative documents that are likely to be considered relevant to a query.

To identify and select this subset of documents, the recommendation is to use more than one IR system and/or different ranking functions. The top n (usually $n = 100$ or 50) documents returned for each retrieval strategy are selected and merged into the pool. Finally, annotators judge only the documents on the pool of each 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

```
<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>
```

Figure 2: 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.

works [4], [3], and [18], have shown that some retrieval quality metrics are robust and stable to handle this scenario.

Our pool was created using two IR systems, Apache Solr[9] and Anserini[10]. In order to select the best configurations, some preliminary experiments were run with the Folha de São Paulo test collection (FSP) [14]. The four configurations (two from each IR system) are presented in Table 2. Having different configurations is important to avoid system bias [13]. 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 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.

3.4 Relevance Assessments

To enable assessing the relevance of the documents w.r.t. the queries, a complete annotation system was developed. 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 1 shows a screenshot of the annotation page.

Table 2: 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	–

To assure the quality of the collection, the relevance judgments were made by annotators with domain knowledge which included geologists and petroleum engineers. 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 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 & Voorhees [4] found that, for mean average precision, 25 topics are the minimum number considered acceptable.

Figure 3 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.

4 EXPERIMENTS

In order to generate baseline results on REGIS that could be used for future research on the topic, we scored the results of the four configurations shown in Table 2 using the relevance assessments. The retrieval quality metrics were calculated using trec_eval[11]. Two scenarios were considered depending on the minimum level of relevance for a document to be considered as relevant. In the first

[9] Solr: https://lucene.apache.org/solr/

[10] Anserini: http://anserini.io/

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

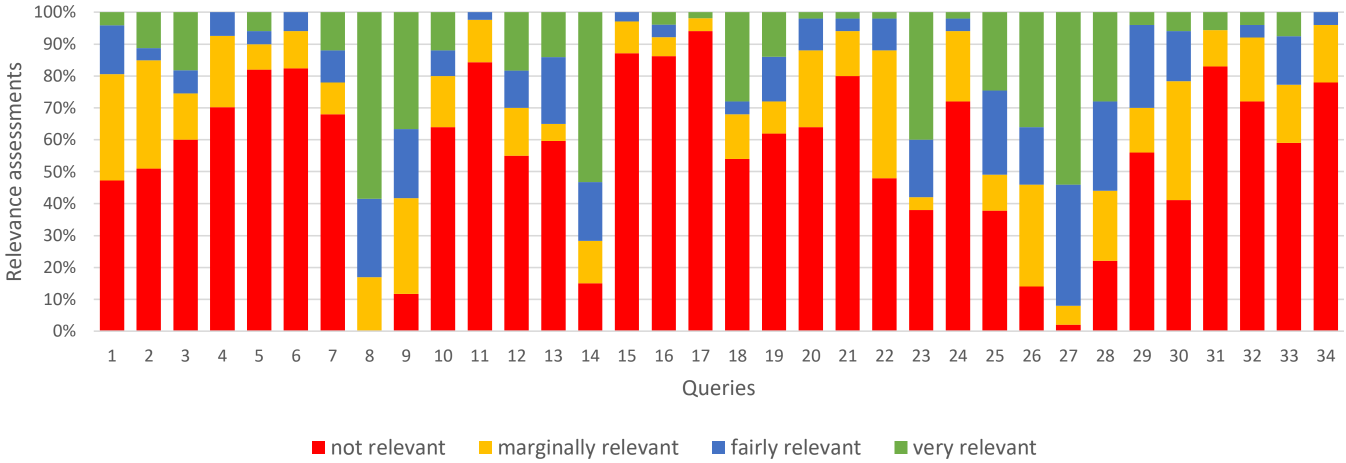


Figure 3: Distribution of the levels of relevance

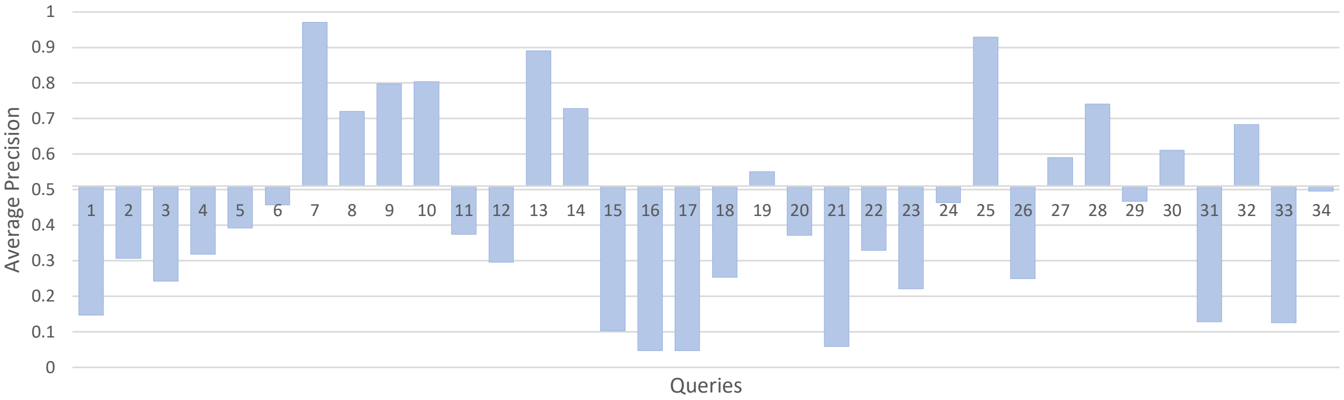


Figure 4: Average Precision by Query for BM25 with proximity under Solr

assessment, the minimum was set to "marginally relevant". Then, in a more strict evaluation, to be considered relevant, the document should be at least "fairly relevant". The results are shown in Table 3. The best configuration was BM25+Prox in both scenarios and in all metrics. The score for NDCG in this configuration is 0.6856. As expected, the metrics are lower when only documents that are at least fairly relevant are considered. In both scenarios, the ranking of the configurations is the same with BM25+RM3 having the worst results. Analyzing the keywords generated by RM3 for query expansion, most of them were not representative words, which can lead the system to retrieve non-relevant documents. Considering the technical terms in the domain and named entities, perhaps a specialized thesaurus would help generate better terms for query expansion.

Figure 4 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. This tendency can be observed comparing the bars in Figures 3 and 4.

Table 3: Baseline Results on REGIS

Minimum Relevance	Scoring Function	MAP	Rel-Ret	PR@10
Marginally Relevant	DFR+Prox	0.4384	553	0.5912
	BM25+Prox	**0.5300**	**633**	**0.6471**
	QLD	0.3462	462	0.4882
	BM25+RM3	0.2746	408	0.4118
Fairly Relevant	DFR+Prox	0.3776	345	0.3491
	BM25+Prox	**0.4747**	**403**	**0.4294**
	QLD	0.2974	301	0.3353
	BM25+RM3	0.2256	261	0.2765

5 DISCUSSION

In this section, we discuss some relevant issues related to the construction of REGIS.

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 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 [13], 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 (Table 3). On the other hand, long documents are less prone to suffer from OCR errors since they have a chance of containing at least one occurrence of an important word that has been extracted correctly.

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.* [2] has shown that if such errors exceed a 5% word error rate, then retrieval quality can be significantly affected. While we did not perform a formal evaluation of the extraction errors in REGIS, we believe they fall below 5% word error rate, as these scanned documents are not the majority.

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 [9, 17]. 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.

6 CONCLUSION AND FUTURE WORK

This paper describes 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. We also carried out experiments with standard retrieval methods that are intended to serve as baseline for future work using REGIS.

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. Also, since the original documents are in PDF, REGIS can be used to test the impact that correcting OCR extraction errors.

ACKNOWLEDGMENTS

This work was partially supported by Petrobras 2017/00752-3, CNPq/Brazil, and by CAPES Finance Code 001. The authors thank Cassiana Michelin, Vinícius Sgorla, Guilherme Bazzo, Givanildo do Nascimento, Lucas Valadares, Antônio Alexandre, Mariana Mendoza, Diogo Magalhães, Delton Vaz, Marcelo Rezende, Fábio Cordeiro, Luciana Bencke, and Rafael da Rocha for their valuable aid in providing relevance judgements. We also thank Gustavo Lorentz and Danny Vargas who collected the geoscientific documents and removed the duplicates.

REFERENCES

[1] Moumita Basu, Anurag Roy, Kripabandhu Ghosh, Somprakash Bandyopadhyay, and Saptarshi Ghosh. 2017. Microblog Retrieval in a Disaster Situation: A New Test Collection for Evaluation.. In *SMERP@ ECIR*. 22–31.

[2] Guilherme Torresan Bazzo, Gustavo Acauan Lorentz, Danny Suarez Vargas, and Viviane P Moreira. 2020. Assessing the Impact of OCR Errors in Information Retrieval. In *European Conference on Information Retrieval*. Springer, 102–109.

[3] Chris Buckley and Ellen M Voorhees. 2004. Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. 25–32.

[4] Chris Buckley and Ellen M Voorhees. 2017. Evaluating evaluation measure stability. In *ACM SIGIR Forum*, Vol. 51. Association for Computing Machinery, New York, NY, USA, 235–242.

[5] Cyril W Cleverdon. 1962. *Aslib Cranfield research project: report on the testing and analysis of an investigation into the comparative efficiency of indexing systems*. Technical Report.

[6] Bernardo Consoli, Joaquim Santos, Diogo Gomes, Fabio Cordeiro, Renata Vieira, and Viviane Moreira. 2020. Embeddings for named entity recognition in geoscience portuguese literature. In *Proceedings of The 12th Language Resources and Evaluation Conference*. 4625–4630.

[7] Fábio Corrêa Cordeiro and Cristian Enrique Munoz Villalobos. 2020. Petrolês - How to Build a Specialized Oil and Gas Corpus in Portuguese. *Rio Oil and Gas Expo and Conference* 20, 2020 (Dec. 2020), 387–388. https://doi.org/10.48072/2525-7579.rog.2020.387

[8] Diogo Gomes, Fábio Cordeiro, Bernardo Consoli, Nikolas Santos, Viviane Moreira, Renata Vieira, Silvia Moraes, and Alexandre Evsukoff. 2021. Portuguese word embeddings for the oil and gas industry: Development and evaluation. *Computers in Industry* 124 (2021), 103347. https://doi.org/10.1016/j.compind.2020.103347

[9] William Hersh, Andrew Turpin, Susan Price, Benjamin Chan, Dale Kramer, Lynetta Sacherek, and Daniel Olson. 2000. 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*. 17–24.

[10] Marianne Lykke, Birger Larsen, Haakon Lund, and Peter Ingwersen. 2010. Developing a test collection for the evaluation of integrated search. In *European Conference on Information Retrieval*. Springer, 627–630.

[11] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Evaluation in information retrieval*. Cambridge University Press, Cambridge, UK, 139–161. https://doi.org/10.1017/CBO9780511809071.009

[12] Anna Ritchie, Simone Teufel, and Stephen Robertson. 2006. Creating a test collection for citation-based IR experiments. In *Proceedings of the human language technology conference of the NAACL, main conference*. 391–398.

[13] Mark Sanderson. 2010. *Test collection based evaluation of information retrieval systems*. Now Publishers Inc.

[14] Diana Santos and Paulo Rocha. 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*. Springer, 821–832.

[15] Ian Soboroff, Kira Griffitt, and Stephanie Strassel. 2016. 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*. 713–716.

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

[17] Andrew Turpin and Falk Scholer. 2006. 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*. 11–18.

[18] Emine Yilmaz and Javed A Aslam. 2006. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM international conference on Information and knowledge management*. 102–111.