Procesamiento del Lenguaje Natural, Revista nº 74, marzo de 2025, pp. 55-65 recibido 18-11-2024 revisado 19-01-2025 aceptado 05-02-2025

Natural Language Processing in Geology: A Review of Portuguese Language Resources and Techniques

Procesamiento del Lenguaje Natural en Geología: Una Revisión de Recursos y Técnicas en Lengua Portuguesa

Rafael Oleques Nunes, Andre Suslik Spritzer, Carla Maria Dal Sasso Freitas, Dennis Giovani Balreira, Joel Luís Carbonera
Universidade Federal do Rio Grande do Sul
{ronunes, spritzer, carla, dgbalreira, jlcarbonera}@inf.ufrgs.br

Abstract: This paper examines Natural Language Processing (NLP) advancements in the geological domain, focusing on Portuguese-language resources and applications. As geological data grows in complexity, there is an increasing need for specialized NLP tools. Recent developments include specialized corpora and adapted models, notably enhancing Named Entity Recognition, Information Retrieval, and dependency parsing. Despite these improvements, gaps remain, particularly in classification tasks and advanced evaluation techniques. This paper addresses these gaps by proposing rethinking methodologies and advocating for further research to broaden geological NLP applications.
Keywords: Survey, resources, corpora, geology.

Resumen: Este artículo examina los avances del Procesamiento del Lenguaje Natural (PLN) en el ámbito geológico, centrándose en los recursos y aplicaciones en lengua portuguesa. A medida que los datos geológicos se vuelven más complejos, existe una creciente necesidad de herramientas de PLN especializadas. Los desarrollos recientes incluyen corpus especializados y modelos adaptados, que mejoran notablemente el reconocimiento de entidades nombradas, la recuperación de información y el análisis de dependencias. A pesar de estas mejoras, aún quedan lagunas, en particular en las tareas de clasificación y las técnicas de evaluación avanzadas. Este artículo aborda estas lagunas proponiendo replantear las metodologías y abogando por una mayor investigación para ampliar las aplicaciones geológicas del PLN.
Palabras clave: Revisión, recursos, corpus, geología.

1 Introduction

The geoscientific domain, especially in industries like oil and gas, generates extensive unstructured data critical for decision-making. These data are usually described in text format, which makes text analysis a crucial method for obtaining valuable information from unstructured textual data (Tveritnev et al., 2023). Traditional Natural Language Processing (NLP) models often struggle with the complexity and specialized nature of geological texts, highlighting the need for domain-specific tools to enhance data extraction, classification, and retrieval.

Recent advancements in NLP have addressed these challenges by developing specialized resources and adapting models to geological contexts (Blondelle and Nguyen-Thuyet, 2023; Holden et al., 2019). These advancements have shown promise in improving the processing of geological information, particularly in the oil and gas industry.

Despite these improvements, significant gaps remain in the literature. Most of the research in the geological domain is devoted to information extraction (Qiu et al., 2023; Qiu et al., 2024; Dong et al., 2023; Wang et al., 2022) and entity recognition (Batista et al., 2010; Sobhana, 2012; Qiu et al., 2019). There is limited research on classification tasks within geological NLP, and advanced evaluation techniques are underutilized. Regarding specific corpus for such tasks, we found works based on Indian (Sobhana, Mitra, and Ghosh, 2010), English (Enkhsaikhan et al., 2021) and Chinese (Qiu et al., 2024).

General research on the field addresses challenges in extracting and structuring knowledge from large volumes of textual data. The study proposed by (Lin et al., 2023) focuses on developing a large-scale

ISSN 1135-5948 DOI 10.26342/2025-74-4

language model specialized in geosciences, adapting AI architectures for this specific domain. Another study investigates named entity recognition in geological texts, using few-shot learning to overcome the scarcity of annotated data (Liu et al., 2022). There are also research efforts aimed at applying NLP to convert geological descriptions into representations usable for predictive modeling (Lawley et al., 2023). Despite of their success, unfortunately none of them focuses on the Portuguese language.

This paper aims to review the current literature in Portuguese language, analyze the advancements, and identify the existing gaps. By highlighting these areas, the study seeks to provide a comprehensive overview and suggest directions for future research in geological NLP.

2 Methodology

Following previous research (Neiva et al., 2016; Jr. et al., 2021), the first step of our process involves determining the guidelines for this systematic review. We aim to select recent papers exploring Portuguese NLP resources or techniques in the geological domain. Next, we detail the rules we followed to obtain the papers analyzed.

The research questions we explore are: (RQ1) What corpora have been developed and used for NLP in the geology domain? (RQ2) How have embeddings and language models been adapted or developed specifically for geology-related tasks? (RQ3) What downstream tasks have been addressed using these specialized resources and models in the geology domain?

We review papers from the past ten years, i.e., from 2015 to 2024. This time range includes the most relevant works, mainly due to the high advances in NLP urged in the last decade. Our study includes only peer-reviewed indexed conference papers and articles written in Portuguese or English. We have not explored other research reports, such as theses, panels, and reports.

We conducted our search on the following academic databases: (i) IEEE Xplore[1], (ii) ACM Digital Library[2], (iii) Springer Link[3], (iv) Scopus[4], (v) Science Direct[5], and (vi) ACL Anthology[6]. The selected papers must be directly related to NLP and geology in Portuguese. Papers may present new resources, techniques, and analyses within this field. We manually investigated the retrieved papers for each database, resulting in a query to check their scope according to their title and abstracts. We excluded any further work which did not fit this scope.

We used two separate queries for ScienceDirect due to the platform's limit of eight connectors per search. The first query was *"natural language processing" AND "portuguese" AND ("geological" OR "geology" OR "geoscience" OR "oil and gas")*, and the second one was *("processamento da linguagem natural" OR "processamento de linguagem natural") AND ("geologia" OR "geológico" OR "óleo e gás" OR "geociência")*. We combined both queries using an OR operator for Springer, IEEE Xplore, ACM Digital Library, Scopus, and ACL Anthology.

3 Review Analysis

Our search retrieved a total of ten papers. This section includes an analysis of these papers according to the research questions we propose.

3.1 Corpora (RQ1)

The development of corpora tailored to the geological domain has been instrumental in advancing various NLP tasks, from Information Retrieval (IR) to Named Entity Recognition (NER) and syntactic parsing. Each corpus is a crucial resource, contributing unique strengths and addressing specific challenges within the field.

REGIS (Reference Corpus for Geoscientific Information Systems) (Lima de Oliveira, Romeu, and Moreira, 2021) is a large-scale corpus focusing on the geoscientific domain, particularly the oil and gas industry. Comprising academic and industry documents written in Portuguese, domain specialists annotate REGIS with relevance judgments ("very relevant", "fairly relevant", "marginally relevant", and "not relevant"). The comprehensive collection process involved filtering documents based on ge-

[1]https://ieeexplore.ieee.org/Xplore/home.jsp
[2]https://dl.acm.org/
[3]https://link.springer.com/
[4]https://www.scopus.com/
[5]https://www.sciencedirect.com/
[6]https://aclanthology.org/

ology and petroleum exploration keywords, with Optical Character Recognition (OCR) used for processing scanned texts. Although REGIS provides a rich resource for IR and other NLP applications, it faces challenges such as the handling of long documents, OCR errors, and syntax variations due to the documents' wide temporal range (60 years).

Moving into syntactic parsing, **PetroGold** (Freitas et al., 2015; de Souza et al., 2021) emerged as a gold standard treebank tailored for the oil and gas domain. Initially consisting of 19 academic documents and 253,640 tokens, it follows the Universal Dependencies (UD) framework for morphosyntactic dependency annotation. PetroGold addresses sector-specific challenges, such as complex nominal structures, bibliographic references, and multiword expressions, further refined in subsequent annotation efforts (de Souza and Freitas, 2023).

The corpus has been expanded in size and scope, integrating additional documents to capture a broader range of petroleum-related content (de Souza and Freitas, 2023). Enhancements in annotation practices involved developing detailed rules and applying advanced machine learning techniques, with a thorough revision process that combined automatic and manual checks to improve accuracy.

Focusing on NER, **PetroNER** (Freitas et al., 2015) is the first published Portuguese corpus specifically tailored to NER's geological domain. Built upon PetroGold, PetroNER was automatically labeled using a set of linguistic rules and supported by the development of the domain ontology **PetroKGraph** (Freitas et al., 2015), which extends the concepts established in **GeoCore** (Garcia et al., 2020). This extensive groundwork enabled the creation of a corpus comprising approximately 20,000 entities, annotated with named entities and part-of-speech tags in the BIO format. PetroGold is the foundation for developing PetroNER annotations (Freitas et al., 2015).

To the best of our knowledge, **GeoCorpus** (Amaral et al., 2017) stands out as the only human-labeled NER corpus specifically designed for Portuguese, focusing on the Brazilian Sedimentary Basin subdomain. This corpus emphasizes the annotation of geological entities within scientific documents, such as theses, dissertations, articles, and technical bulletins. The semi-manual annotation process, facilitated by the IdENGeo tool and carried out by geology experts, ensures that the entities are highly relevant to the Brazilian geological context. GeoCorpus has been further enriched through subsequent updates, resulting in **GeoCorpus-2** (Consoli et al., 2020) and **GeoCorpus-3** (Gomes et al., 2021), addressing issues such as duplicate sentences, format conversions, and refined annotations.

Relevance and Impact. These corpora enable tasks like IR, NER, and syntactic parsing in areas where specialized knowledge is essential. The annotation processes improve the accuracy of language models and set new benchmarks for creating resources tailored to specific domains. These works help to narrow the gap between general NLP models and the unique requirements of the geological and petroleum industries, where specialized understanding is vital.

Literature Gap. While these corpora have made significant strides, there is still a notable gap in the availability of resources across different geological subdomains. Existing corpora like PetroNER and GeoCorpus mainly focus on the oil and gas industry or specific areas like the Brazilian Sedimentary Basin. However, there is a shortage of comprehensive corpora that cover a wider range of geological topics, especially in Portuguese. Other important NLP tasks, such as text classification and summarization, lack the specialized corpora needed for effective application in the geological domain. The current NLP tools are still in the early stages of adapting to the unique challenges of geological texts, such as complex syntax and specialized terminology.

Future efforts should aim to create more diverse and multidisciplinary corpora and improve the tools to better handle the intricacies of geological language. Developing resources that span a broader range of geological subdomains and NLP tasks will be crucial for advancing the field and enabling more sophisticated applications of NLP in geology.

3.2 Embeddings (RQ2)

Several studies have explored developing and applying specialized word embeddings (WE) tailored for geological terminology (Gomes, Cordeiro, and Evsukoff, 2018; Consoli et al., 2020; Gomes et al., 2021). Techniques

like Word2Vec (Mikolov, 2013) and FastText (Bojanowski et al., 2017) have generated embeddings that capture the intricate semantic relationships and nuances of domain-specific vocabulary. These embeddings are crucial for enhancing the understanding of geological terms by encoding their contextual meanings within a specialized corpus.

In the first study to create Portuguese WE specialized in the geological domain (Gomes, Cordeiro, and Evsukoff, 2018), the authors used Word2Vec (Mikolov, 2013), FastText (Bojanowski et al., 2017), and GloVe (Pennington, Socher, and Manning, 2014) models. They generated embeddings from a specialized corpus that included Petrobras' Technical Bulletins, Geotechnical Bulletins, and other geological texts from the Brazilian National Agency of Petroleum, Natural Gas, and Biofuels (ANP). This approach effectively captured domain-specific terminology, making the embeddings well-suited to the unique language of the geological field. Table 1 shows this study's document distribution.

To evaluate the effectiveness of these embeddings, the study conducted a qualitative analysis against a set of geological vocabulary, comparing the specialized embeddings with the NILC embeddings, which were trained on a general Portuguese corpus (Hartmann et al., 2017). The analysis highlighted the superior ability of the specialized embeddings to capture domain-specific nuances, thereby providing valuable resources for geosciences and facilitating more accurate and context-aware NLP applications in Portuguese. As a foundational study, it paved the way for subsequent advances in geological NLP.

Different embedding architectures were explored and evaluated against existing Portuguese models (Consoli et al., 2020), including those developed in earlier research (Gomes, Cordeiro, and Evsukoff, 2018). While the main focus of this study was on enhancing NER performance, rather than solely on the embeddings themselves, it emphasized the importance of tailoring embeddings to specific domains. The discussion on the impact of these domain-specific models in NER will be expanded in Section 3.3.

A subsequent study established the current state-of-the-art (SOTA) in geological embeddings for Portuguese (Gomes et al., 2021). This research highlighted that embeddings trained specifically on geological texts significantly outperform those trained on general-domain data for geological NLP tasks. The study involved two WE models: one trained solely on geological texts (refer to the *specific* row in Table 2) and another trained on a mix of geological and general texts (refer to the *hybrid* row in Table 2). Both models demonstrated SOTA performance in intrinsic evaluations, such as word similarity tasks, and extrinsic evaluations, such as NER.

Regarding contextual embeddings, a study introduced PetroBERT, the first Bidirectional Encoder Representations for Transformers (BERT) model (Devlin et al., 2019) specifically designed for the geological domain (Rodrigues et al., 2022). This model builds on the multilingual BERT (mBERT) (Devlin et al., 2019) and BERTimbau (Souza, Nogueira, and Lotufo, 2020) by further pretraining them on geological texts. The training corpus for PetroBERT included datasets from the SOTA word embeddings approach (Gomes et al., 2021), along with a proprietary Daily Drilling Reports corpus (DDR-Corpus) containing 29,000 sentences. Since the evaluation of PetroBERT mainly focused on downstream tasks, further details will be discussed in Section 3.3.

Relevance and Impact. These studies collectively underscore the crucial role of domain-specific NLP resources and models in advancing research and applications within the geological sector, particularly in the oil and gas industry. By developing and deploying these specialized resources, researchers and practitioners can achieve more precise and efficient processing of geological texts than possible with general-purpose Portuguese models. This advancement accelerates research efforts and enhances decision-making processes in geological applications, demonstrating the significant impact of tailored NLP solutions in this field.

Literature Gap. Despite the advancements made, there remains a significant gap in applying recent transformer architectures, such as Llama and T5, within the geological NLP context, along with the need to explore further adapting encoder-only models (like BERT) to the geological domain. Developing multilingual models is also crucial to facilitate knowledge transfer across differ-

Institution	Base	Documents / Terms	Tokens (raw / final)	Vocabulary (raw / final)
Petrobras	Technical Bulletins: - Geosciences Bulletins - Petrobras Technical Bulletins - Oil Production Bulletins	∼2000 articles	10,229,664 / 4,962,545	829,907 / 86,862
Petrobras and ANP	Petrobras Glossary, ANP Glossary, Oil Dictionary-Sigla	562 terms, 1255 terms	47,470 / 30,680	12,215 / 7,334
ANP	PRH Final Reports	316 documents	6,288,925 / 3,308,466	394,771 / 52,880
ANP	Publications	168 documents	2,415,849 / 1,138,685	115,345 / 18,236
ANP	Technical Notes and Studies	141 documents	1,238,429 / 669,356	91,391 / 16,922
Total			**20,220,337 / 10,109,732**	**113,934**

Table 1: Summary of Data by Institution (Gomes, Cordeiro, and Evsukoff, 2018).

Corpus Domain	Source	Description	Sentences	Tokens
Specific	Petrobras	Bulletins of Geosciences and Petroleum Production	298,865	3,821,966
		Theses and dissertations on the O&G domain	2,939,262	37,024,438
	ANP	Bulletins and technical reports	132,955	2,136,465
		Theses and dissertations on the O&G domain	279,196	3,629,999
	IBP	Proc. of the Rio O&G Conf.	89,116	1,287,223
	IBICT-BDTD	Brazilian database of theses and dissertations, filtered by petroleum-related domains	2,558,837	37,825,743
	Total		**6,295,231**	**85,725,834**
Generic	Hartmann et al. (Hartmann et al., 2017)	The public part of a general-domain corpus in Portuguese	37,327,741	365,295,169
Hybrid		A combination of the domain-specific and general-domain corpora	43,622,972	451,021,003

Table 2: Sumarry of the Petroles corpus (Gomes et al., 2021).

ent languages. Another promising area for exploration is multimodal learning, which integrates textual and visual data (e.g., geological maps and core sample images) to enhance NLP performance further.

3.3 Downstream Tasks (RQ3)

Geological NLP research has mainly concentrated on Named Entity Recognition, with several studies evaluating its effectiveness on open-domain corpora (Consoli et al., 2020;

Gomes et al., 2021; Rodrigues et al., 2022). Information Retrieval has been explored to a lesser degree (Lima de Oliveira, Romeu, and Moreira, 2021), while dependency parsing has shown potential in specialized corpora like PetroGold (Freitas et al., 2015; de Souza et al., 2021). Classification tasks remain underexplored, with only one study addressing this area (Rodrigues et al., 2022). This section summarizes the progress in NER and IR, noting the limited focus on dependency parsing and classification.

3.3.1 Named Entity Recognition

In the context of **NER**, the current literature reveals three main approaches: traditional NLP models (Amaral, 2017), bidirectional long short-term memory with a conditional random field (BiLSTM-CRF) models (Lample et al., 2016) with domain-specific WE (Consoli et al., 2020; Gomes et al., 2021), and transformer-based models (Rodrigues et al., 2022). These approaches have been applied to variations of the GeoCorpus dataset (Amaral, 2017; Consoli et al., 2020; Gomes et al., 2021), as shown in Table 3. Notably, no studies have utilized the PetroNER corpus (Freitas et al., 2015) for training classifiers.

Traditional models such as Naive Bayes (Maron, 1961), J48 Decision Tree (Quinlan, 2014), and CRF (Lafferty et al., 2001) were explored in the GeoCorpus monograph (Amaral, 2017). Although these models achieved lower results, with CRF attaining the highest performance at **54.33%**, they provide crucial baselines for evaluating advanced approaches.

The study by Consoli et al. (Consoli et al., 2020) introduced WE with BiLSTM-CRF to the 13 classes of GeoCorpus-2. It found that domain-specific WE (GeoWE) (Gomes, Cordeiro, and Evsukoff, 2018) underperformed compared to general Portuguese WE (Hartmann et al., 2017), with micro F1-Scores of 53.71% and 71.33%, respectively. Interestingly, traditional models like CRF achieved results comparable to those of domain-specific WE. The study also explored stacked embeddings and various configurations, achieving a notable improvement with a micro F1-Score of **84.63%** using the best combination of embeddings and models.

In a subsequent study, new WE tailored for the oil and gas industry were developed (Gomes et al., 2021). This research demonstrated that domain-specific embeddings from the Petroles corpus (Freitas et al., 2015), when used with a BiLSTM-CRF model, outperformed previous approaches, achieving a micro F1-Score of **86.00%** on GeoCorpus-3. Unlike GeoCorpus-2, which focuses on 13 classes, GeoCorpus-3 emphasizes the top ten classes with the most information. The enhanced performance can be attributed to the more detailed and specialized annotations of GeoCorpus-3.

Despite these advancements with domain-specific embeddings, PetroBERT (Rodrigues et al., 2022), a BERT-based model tailored for the geological domain, achieved relatively lower results compared to earlier approaches using GeoCorpus-2. Although PetroBERT builds on multilingual BERT and BERTimbau and was trained on a range of geological texts, its performance did not surpass that of specialized embeddings. Notably, while multilingual BERT and BERTimbau also yielded lower results compared to domain-specific embeddings, they outperformed PetroBERT. Specifically, BERTimbau achieved a best result of 82.88%, whereas PetroBERT only reached 77.28%. This underscores that, despite the sophisticated architecture of BERT-based models, they may not always outperform well-optimized domain-specific embeddings and highlights the need for refinement and adaptation to specific tasks and datasets.

Recently, Nunes et al. (Nunes et al., 2024) conducted a comprehensive evaluation of transformer-based language models in the context of NER for geosciences. Using the latest version of GeoCorpus (GeoCorpus-3), the study compared the performance of BERTimbau and XLM-RoBERTa (Conneau et al., 2020) with different classification layers (linear and CRF), assessing their effectiveness across 30 geological entity classes. The results indicated that transformer-based models, particularly XLM-RoBERTa with CRF, outperform previous approaches, achieving an F1-Score of up to **89.13%**. In addition to performance analysis, the study employed statistical tests to validate significant differences between models and publicly releases the updated corpus and trained models, fostering advancements in NER research for specialized domains. This evaluation highlights the importance of adapting NLP models to technical fields, emphasizing that architectural choices can significantly impact final performance.

Model Information		Corpus (#Classes)			
Type	**Model**	**v1 (13)**	**v2 (13)**	**v3 (top 10)**	**v3 (30)**
Traditional	J48 Decision Tree	25.50 [R1]			
	Naive Bayes	49.47 [R1]			
	CRF	54.33 [R1]			
Word Embeddings	GeoWE		53.71 [R2]	78.00 [R3]	
	skipgram-NILC		71.33 [R2]	83.00 [R3]	
	PetroVec-O&G			86.00 [R3]	
	PetroVec-hybrid			86.00 [R3]	
Flair Embeddings	FlairBBP		83.10 [R2]		
	FlairBBPGeoFT		84.20 [R2]		
Stacked Embeddings	GeoWE+FlairBBP		78.84 [R2]		
	W2V-SKPG+FlairBBP		84.04 [R2]		
	GeoWE+FlairBBPGeoFT		83.74 [R2]		
	W2V-SKPG+FlairBBPGeoFT		84.63 [R2]		
Transformers	PetroBERT (Linear)		77.28 [R4]		
	mBERT (Linear)		81.63 [R4]		
	BERTimbau (Linear)		82.88 [R4]		87.06 [R5]
	BERTimbau (CRF)				88.93 [R5]
	XLM-RoBERTa (Linear)				88.78 [R5]
	XLM-RoBERTa (CRF)				**89.13 [R5]**

Table 3: NER performance in geology across different studies. Each value corresponds to the micro F1-score reported in the respective reference: [R1] (Amaral, 2017), [R2] (Consoli et al., 2020), [R3] (Gomes et al., 2021), [R4] (Rodrigues et al., 2022), and [R5] (Nunes et al., 2024). The datasets used are versions of GeoCorpus (v1, v2, and v3), representing successive refinements and expansions of the original corpus. This table is an adapted version of the one presented in Nunes et al. (Nunes et al., 2024).

3.3.2 Information Retrieval

Regarding **IR**, The REGIS corpus (Lima de Oliveira, Romeu, and Moreira, 2021) has been utilized with key models, including Best Matching 25 (BM25) (Robertson and Walker, 1994) and Divergence From Randomness (DFR) (Amati and Van Rijsbergen, 2002), combined with techniques like query expansion using Relevance Model 3 (RM3) (Lavrenko and Croft, 2017) and Query Likelihood with Dirichlet smoothing (QLD) (Zhai and Lafferty, 2004). The evaluation of these models involved measuring effectiveness using standard metrics such as Mean Average Precision (MAP), Precision at Rank 10 (PR@10), and Normalized Discounted Cumulative Gain (NDCG). The BM25 model with proximity search achieved the best performance (see Table 4), emphasizing the importance of term proximity in long, technical documents typical of the geoscientific domain.

These results highlight the relevance of REGIS for downstream tasks such as improving domain-specific search systems, automatic query expansion, and assessing OCR impact on IR performance, particularly in underrepresented languages like Portuguese.

3.3.3 Dependency Parsing

Another significant task in NLP is **dependency parsing task**, which involves identifying grammatical relationships between words in a sentence and mapping these relationships into a syntactic structure. PetroGold (Freitas et al., 2015; de Souza et al., 2021) has been utilized to train and evaluate syntactic parsers using tools like UDPipe (Straka, Hajic, and Straková, 2016) and Stanza (Qi et al., 2020), achieving a 90.65% Unlabeled Attachment Score (UAS) and 88.53% Labeled Attachment Score (LAS) (de Souza et al., 2021). The latest PetroGold version, evaluated with UDPipe models, reached notable scores of 98.63% for Universal POS tagging (UPOS), 90.22% for LAS, and 85.61% for Content-Word LAS (CLAS).

3.3.4 Classification

The unique paper addressing **classification** tasks (Rodrigues et al., 2022) uses the DDR corpus. This corpus includes over 29,000 sentences describing the drilling process for 302 wells, with sentences categorized into three levels—activity, operation, and step—reflecting the granularity of events during drilling.

The domain-adapted PetroBERT model outperformed the general-purpose mBERT

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

Table 4: Retrieval performance on the REGIS corpus, as reported in the original study (Lima de Oliveira, Romeu, and Moreira, 2021). This table is directly reproduced from that work, as it remains the only published evaluation of this dataset.

and BERTimbau in sentence classification within the DDR corpus, especially when fine-tuned with domain-specific texts. The best-performing model achieved an F1-score of **51.98%**, illustrating the advantages of domain-specific pre-training for classification tasks in specialized fields. In contrast, BERTimbau achieved a comparable result of 48.24%. This suggests the need for a more detailed analysis, such as cross-validation, to obtain average performance, standard deviation, and statistical significance to better understand the robustness and reliability of the results.

Relevance and Impact. Research on downstream tasks in geological NLP is critically important, with broad relevance across various industrial applications. Advances in NER and IR models, particularly those tailored to the geological domain, significantly enhance data extraction and search functionalities within geoscientific contexts.

The improved performance of domain-adapted models, such as those designed for oil and gas exploration, not only pushes the boundaries of NLP in geology but also delivers practical benefits to industries that heavily rely on accurate geological data. Enhanced NER and IR capabilities allow for more precise and efficient retrieval of geoscientific information, supporting crucial decision-making processes in areas like exploration, drilling, and resource management. Moreover, insights from this research can inform the development of more robust NLP tools for other specialized domains, thereby amplifying the overall impact of this work across multiple fields.

Literature Gap. Despite the significant strides made in NER and IR for geological NLP, several key literature gaps persist. One major gap is the limited exploration of classification tasks within the geological domain. The study on PetroBERT (Rodrigues et al., 2022) stands as one of the few that addresses this aspect, underscoring the need for more research focused on classification methods and their application to geological texts. Expanding this study area, particularly through open corpora, would enhance reproducibility and enable broader evaluation across the research community.

Additionally, there is a noticeable lack of comprehensive research employing advanced evaluation techniques, such as cross-validation and statistical testing, to rigorously assess the reliability and consistency of NLP models in geological contexts. Addressing these gaps could provide a more nuanced understanding of the effectiveness of different NLP approaches in the geological domain, driving further advancements and more reliable applications in the field.

4 Concluding Remarks

Advances in Natural Language Processing for the geoscientific domain, particularly in Portuguese, have significantly improved tasks such as Named Entity Recognition, Information Retrieval, Dependency Parsing, and Classification. Developing domain-specific models and corpora enhances the accuracy and efficiency of processing geological data, benefiting industries like oil and gas exploration.

These advancements have practical implications for industries relying on precise geoscientific information, supporting better decision-making and operational efficiency. However, gaps still need to be addressed, including limited research on classification tasks and the need for advanced evaluation techniques. Future work should focus on expanding research to include more subdomains

and tasks and use open corpora to improve reproducibility.

In summary, while progress has been significant, addressing these gaps is essential for further advancements. Continued development of NLP tools tailored to geological texts will drive innovation and application across the geoscientific and industrial sectors.

Acknowledgements

This work has been partially funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. We also acknowledge the financial support from the Brazilian funding agency CNPq and Petrobras.

References

Amaral, D. O. F. 2017. *Reconhecimento de entidades nomeadas na área da geologia: bacias sedimentares brasileiras*. Pontifícia Universidade Católica do Rio Grande do Sul (Tese de Doutorado).

Amaral, D. O. F., S. Collovini, A. Figueira, R. Vieira, and M. Gonzalez. 2017. Processo de construçao de um corpus anotado com entidades geológicas visando ren. In *Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 63–72. SBC.

Amati, G. and C. J. Van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389.

Batista, D., M. J. Silva, F. Couto, and B. Behera. 2010. Geographic signatures for semantic retrieval. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, pages 18–19.

Blondelle, H. and O. Nguyen-Thuyet. 2023. Natural language processing for key geology information detection and geological description classification. 2023(1):1–5.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 06.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Consoli, B., J. Santos, D. Gomes, F. Cordeiro, R. Vieira, and V. Moreira. 2020. Embeddings for named entity recognition in geoscience portuguese literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4625–4630.

de Souza, E. and C. Freitas. 2023. Explorando variaçoes no tagset e na anotaçao universal dependencies (ud) para português: Possibilidades e resultados com base no treebank petrogold. In *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 125–134. SBC.

de Souza, E., A. Silveira, T. Cavalcanti, M. C. Castro, and C. Freitas. 2021. Petrogold–corpus padrão ouro para o domínio do petróleo. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 29–38. SBC.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dong, J., Q. Qiu, Z. Xie, K. Ma, A. Hu, and H. Wang. 2023. Understanding table content for mineral exploration reports using deep learning and natural language processing. *Ore Geology Reviews*, 156:105383.

Enkhsaikhan, M., W. Liu, E.-J. Holden, and P. Duuring. 2021. Auto-labelling entities in low-resource text: a geological case study. *Knowledge and Information Systems*, 63(3):695–715.

Freitas, C., E. Souza, M. C. Castro, T. Cavalcanti, P. F. da Silva, and F. C. Cordeiro.

2015. Recursos linguísticos para o pln específico de domínio: o petrolês. *Linguamática*, 15(2):51–68.

Garcia, L. F., M. Abel, M. Perrin, and R. dos Santos Alvarenga. 2020. The geocore ontology: a core ontology for general use in geology. *Computers & Geosciences*, 135:104387.

Gomes, D., F. Cordeiro, and A. Evsukoff. 2018. Word embeddings em português para o domínio específico de óleo e gás. In *Proceedings of the 19th Rio oil & gas expo and conference*, page 10.

Gomes, D. d. S. M., F. C. Cordeiro, B. S. Consoli, N. L. Santos, V. P. Moreira, R. Vieira, S. Moraes, and A. G. Evsukoff. 2021. Portuguese word embeddings for the oil and gas industry: Development and evaluation. *Computers in Industry*, 124:103347.

Hartmann, N., E. Fonseca, C. Shulby, M. Treviso, J. Rodrigues, and S. Aluisio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*.

Holden, E.-J., W. Liu, T. Horrocks, R. Wang, D. Wedge, P. Duuring, and T. Beardsmore. 2019. Geodoca – fast analysis of geological content in mineral exploration reports: A text mining approach. *Ore Geology Reviews*, 111:102919.

Jr., J. G., R. Mello, A. Reis, V. Ströele, and J. Souza. 2021. Uma revisão breve sobre perguntas complexas em bases de conhecimento para sistemas de perguntas e respostas. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 283–294, Porto Alegre, RS, Brasil. SBC.

Lafferty, J., A. McCallum, F. Pereira, et al. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*, volume 1, page 3. Williamstown, MA.

Lample, G., M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. 2016. Neural architectures for named entity recognition. In K. Knight, A. Nenkova, and O. Rambow, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June. Association for Computational Linguistics.

Lavrenko, V. and W. B. Croft. 2017. Relevance-based language models. In *ACM SIGIR Forum*, volume 51, pages 260–267. ACM New York, NY, USA.

Lawley, C. J., M. G. Gadd, M. Parsa, G. W. Lederer, G. E. Graham, and A. Ford. 2023. Applications of natural language processing to geoscience text data and prospectivity modeling. *Natural Resources Research*, 32(4):1503–1527.

Lima de Oliveira, L., R. K. Romeu, and V. P. 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*, pages 2363–2368.

Lin, Z., C. Deng, L. Zhou, T. Zhang, Y. Xu, Y. Xu, Z. He, Y. Shi, B. Dai, Y. Song, et al. 2023. Geogalactica: A scientific large language model in geoscience. *arXiv preprint arXiv:2401.00434*.

Liu, H., Q. Qiu, L. Wu, W. Li, B. Wang, and Y. Zhou. 2022. Few-shot learning for name entity recognition in geological text based on geobert. *Earth Science Informatics*, 15(2):979–991.

Maron, M. E. 1961. Automatic indexing: an experimental inquiry. *Journal of the ACM (JACM)*, 8(3):404–417.

Mikolov, T. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Neiva, F. W., J. M. N. David, R. Braga, and F. Campos. 2016. Towards pragmatic interoperability to support collaboration: A systematic review and mapping of the literature. *Information and Software Technology*, 72:137–150.

Nunes, R. O., A. S. Spritzer, D. G. Balreira, C. M. Freitas, and J. L. Carbonera. 2024. An evaluation of large language models for geological named entity recognition. In *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 494–501. IEEE.

Pennington, J., R. Socher, and C. Manning. 2014. GloVe: Global vectors for word representation. In A. Moschitti, B. Pang, and W. Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Qi, P., Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Qiu, Q., Z. Xie, L. Wu, L. Tao, and W. Li. 2019. Bilstm-crf for geological named entity recognition from the geoscience literature. *Earth Science Informatics*, 12:565–579.

Qiu, Q., Y. Duan, K. Ma, L. Tao, and Z. Xie. 2023. Information extraction and knowledge linkage of geological profiles and related contextual texts from mineral exploration reports for geological knowledge graphs construction. *Ore Geology Reviews*, 163:105739.

Qiu, Q., M. Tian, L. Tao, Z. Xie, and K. Ma. 2024. Semantic information extraction and search of mineral exploration data using text mining and deep learning methods. *Ore Geology Reviews*, 165:105863.

Quinlan, J. R. 2014. *C4. 5: programs for machine learning*. Elsevier.

Robertson, S. E. and S. Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 232–241. Springer.

Rodrigues, R. B., P. I. Privatto, G. J. de Sousa, R. P. Murari, L. C. Afonso, J. P. Papa, D. C. Pedronette, I. R. Guilherme, S. R. Perrout, and A. F. Riente. 2022. Petrobert: a domain adaptation language model for oil and gas applications in portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 101–109. Springer.

Sobhana, N. 2012. Enhancing retrieval of geological text using named entity disambiguation. *International Journal of Emerging Technology and Advanced Engineering*, 2:2250–2459, January.

Sobhana, N., P. Mitra, and S. Ghosh. 2010. Conditional random field based named entity recognition in geological text. *International Journal of Computer Applications*, 1:143–147, February.

Souza, F., R. Nogueira, and R. Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer.

Straka, M., J. Hajic, and J. Straková. 2016. Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

Tveritnev, A., M. Khanji, S. Abdullah, L. Rojas, A. Ermilov, F. Al Mansoori, and A. Alblooshi. 2023. Applying machine learning nlp algorithm for reconciliation geology and petrophysics in rock typing. Day 2 Tue, October 03, 2023:D021S054R001, 10.

Wang, B., K. Ma, L. Wu, Q. Qiu, Z. Xie, and L. Tao. 2022. Visual analytics and information extraction of geological content for text-based mineral exploration reports. *Ore Geology Reviews*, 144:104818.

Zhai, C. and J. Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214.