Computers in Industry 124 (2021) 103347

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.elsevier.com/locate/compind

Portuguese word embeddings for the oil and gas industry: Development and evaluation

Diogo da Silva Magalhães Gomes [a,b,*], Fábio Corrêa Cordeiro [a,c], Bernardo Scapini Consoli [d], Nikolas Lacerda Santos [d], Viviane Pereira Moreira [e], Renata Vieira [d,f], Silvia Moraes [d], Alexandre Gonçalves Evsukoff [b]

[a] Petrobras Research and Development Center (CENPES), Rio de Janeiro, Brazil
[b] Federal University of Rio de Janeiro (COPPE/UFRJ), Rio de Janeiro, Brazil
[c] Getulio Vargas Foundation - School of Applied Mathematics (FGV), Rio de Janeiro, Brazil
[d] Pontifical Catholic University of Rio Grande do Sul (PUC-RS), Rio Grande do Sul, Brazil
[e] Federal University of Rio Grande do Sul (UFRGS), Rio Grande do Sul, Brazil
[f] University of Evora, CIDEHUS (UE-PT), Évora, Portugal

ARTICLE INFO

Article history:
Received 2 June 2020
Received in revised form 3 November 2020
Accepted 5 November 2020
Available online 26 November 2020

Keywords:
NLP
Machine learning
Oil and gas
Word embeddings

ABSTRACT

Over the last decades, oil and gas companies have been facing a continuous increase of data collected in unstructured textual format. New disruptive technologies, such as natural language processing and machine learning, present an unprecedented opportunity to extract a wealth of valuable information within these documents. Word embedding models are one of the most fundamental units of natural language processing, enabling machine learning algorithms to achieve great generalization capabilities by providing meaningful representations of words, being able to capture syntactic and semantic features based on their context. However, the oil and gas domain-specific vocabulary represents a challenge to those algorithms, in which words may assume a completely different meaning from a common understanding. The Brazilian pre-salt is an important exploratory frontier for the oil and gas industry, with increasing attractiveness for international investments in exploration and production projects, and most of its documentation is in Portuguese. Moreover, Portuguese is one of the largest languages in terms of number of native speakers. Nonetheless, despite the importance of the petroleum sector of Portuguese speaking countries, specialized public corpora in this domain are scarce. This work proposes `PetroVec`, a representative set of word embedding models for the specific domain of oil and gas in Portuguese. We gathered an extensive collection of domain-related documents from leading institutions to build a large specialized oil and gas corpus in Portuguese, comprising more than 85 million tokens. To provide an intrinsic evaluation, assessing how well the models can encode domain semantics from the text, we created a semantic relatedness test set, comprising 1,500 word pairs labeled by selected experts in geoscience and petroleum engineering from both academia and industry. In addition, we performed an extrinsic quantitative evaluation on a downstream task of named entity recognition in geoscience, plus a set of qualitative analyses, and conducted a comparative evaluation against a public general-domain embedding model. The obtained results suggest that our domain-specific models outperformed the general model on their ability to represent specialized terminology. To the best of our knowledge, this is the first attempt to generate and evaluate word embedding models for the oil and gas domain in Portuguese. Finally, all the resources developed by this work are made available for public use, including the pre-trained specialized models, corpora, and validation datasets.

1. Introduction

Over the last decades, companies have gathered huge amounts of data stored in unstructured textual format. Considerable potentially valuable information may be hidden within these increasing volumes of documents such as scientific articles, journals, technical reports, operation logs and laboratory analysis. Having all those textual resources correctly identified and processed is crucial to support a wide range of industrial and academic applications (Ittoo et al., 2016). Considering the intense competitiveness in this industrial environment, it has been economically vital for oil and gas (O&G) companies to fully leverage information from their exist-

* Corresponding author.
E-mail address: diogosmg@ufrj.br (D.d.S.M. Gomes).

https://doi.org/10.1016/j.compind.2020.103347
0166-3615/

ing data sources, in order to accelerate the pursuit for maximizing their operational efficiency (Blinston and Blondelle, 2017; Lu et al., 2019).

Some recent advances in natural language processing (NLP) with deep learning algorithms (LeCun et al., 2015; Goodfellow et al., 2016) were successfully applied by several industrial applications, providing efficiency improvements in their decision-making processes (Ittoo et al., 2016; Blinston and Blondelle, 2017; Young et al., 2018; Nooralahzadeh et al., 2018; Cordeiro et al., 2019). Those algorithms take unstructured text as their basic input, therefore it is important to obtain suitable mathematical representations for the textual elements. Word embedding models have been efficiently used to provide such meaningful representations, which consist of applying unsupervised learning methods on a text corpus to assign a dense n-dimensional vector to each word in a vocabulary. These models can encode semantic and syntactic similarities between words based on the context where they occur (Mikolov et al., 2013a, 2013; Hartmann et al., 2017). These word vector representations are one of the most fundamental units in any NLP application, since they allow machine learning algorithms to achieve better accuracy due to their great generalization capability (Goldberg, 2016).

Several general-domain embeddings for different languages are available for public use (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2017; Fares et al., 2017), including a few models for Portuguese (Rodrigues et al., 2016; Hartmann et al., 2017). However, the highly technical vocabulary of the O&G domain presents a challenge to NLP applications, since some words may assume a completely different meaning from their conventional interpretation (Nooralahzadeh et al., 2018; Cordeiro et al., 2019). For example, a "*christmas tree*" is an assembly of valves that provides flow control on a oil or gas well – a vector representation drawn from general-domain corpora would hardly capture the intended meaning. Furthermore, there is consistent evidence that developing specialized word embedding models induced from a domain-specific corpus can significantly improve the quality of their semantic representation (Lai et al., 2016; Pakhomov et al., 2016; Nooralahzadeh et al., 2018; Wang et al., 2018a; Alsentzer et al., 2019; Tshitoyan et al., 2019).

Portuguese is one of the languages with the largest number of native speakers. Moreover, recent auction offers for Brazilian pre-salt exploration blocks and improvements on regulatory frameworks have increased the attractiveness for international investments in exploration and production projects (Clavijo et al., 2019). But, despite the importance of the petroleum sector of Portuguese speaking countries, specialized public corpora in this domain are scarce. Furthermore, technical texts in the O&G domain have known differences in linguistics properties and meanings that differ from general-domain texts, motivating the need for specialized embeddings representations for NLP tasks.

Aiming at filling this gap, we introduce `PetroVec`, a set of specialized pre-trained word embedding models for the O&G domain in Portuguese. `PetroVec` was trained on a large O&G corpus, which we assembled using thousands of documents such as periodicals, technical reports, glossaries, academic theses, and articles, published by both academia and major companies. We trained the word embedding models from the specialized corpus using Word2vec (Mikolov et al., 2013a) and FastText (Bojanowski et al., 2017), exploring some variations of corpora composition. Since there is a lack of resources to evaluate word embedding models on this domain and language, we created a test set containing semantic relatedness scores for 1500 word pairs, labeled by experts in geosciences and petroleum engineering from both academia and industry. Hence, we were able to perform intrinsic evaluations, assigning a metric of how well the embeddings can encode semantic properties from the corpus. Additionally, we also performed extrinsic evaluations on a downstream task of named entity recognition in geoscience, plus a set of qualitative analyses. With that, our models were thoroughly evaluated, both quantitatively (with intrinsic and extrinsic evaluations) and qualitatively. Furthermore, we conducted a comprehensive analysis comparing our models and a pre-trained general-domain model in Portuguese (Hartmann et al., 2017). Our findings confirm that our specific-domain models capture semantics in a way that is closer to domain experts, with all evaluation alternatives pointing to the same conclusions.

Finally, all the resources developed in this work are available for public use in our repository[2], including the pre-trained word embeddings, corpora, test sets and scripts for preprocessing, training and evaluating the models. The main contributions of this work are as follows: (i) a representative set of domain-specific word embedding models for the O&G industry in Portuguese; (ii) the largest corpus ever reported for this domain and language; and (iii) the first annotated test set for intrinsic semantic evaluation for the O&G domain in Portuguese. We believe that many researchers working in Portuguese O&G domain related projects, both from the industry and academia, can benefit from these resources.

The remainder of this article is organized as follows: Section 2 introduces the background concepts. Section 3 surveys the related work in domain-specific word embeddings. In Section 4, we describe the corpus assembly and the training of the embeddings. Then, the next sections report on the different evaluations we performed. Section 5 presents the intrinsic evaluation. The extrinsic evaluation is detailed in Section 6. Section 7 presents the qualitative evaluation. Finally, Section 8 concludes the article.

2. Background

Natural language processing (NLP) encompasses a set of computational techniques which aim to provide algorithms the ability to automatically analyze text written in human language. These techniques aim to resolve syntactic structure, word disambiguation, and comprehend the semantic scope of a sentence (Manning and Schutze, 1999). Such algorithms have been successfully applied to many downstream tasks, both in academia and industry, such as automatic machine translation (Vaswani et al., 2017; Ruder et al., 2019), named entity recognition (Yadav and Bethard, 2018), text classification (Kowsari et al., 2019), sentiment analysis (Zhang et al., 2018), question answering (Young et al., 2018), information extraction (Niklaus et al., 2018), semantic search (Bast et al., 2016), and automatic text summarization (Allahyari et al., 2017). Over the last decades, computational linguistics has made an impressive progress, especially due to the large availability of textual resources, advances on human language research, improvements on computational power and the development of new successful machine learning (ML) methods (Hirschberg and Manning, 2015). These sophisticated ML approaches are capable of processing large amounts of data efficiently, having evolved from shallow neural network models based on sparse vectors to deep learning methods consuming dense vector representations (Young et al., 2018). One of the main contributions to the success of these models may be attributed to the advent of distributed vector representations and their high generalization capabilities (Manning, 2015).

2.1. Word embeddings

Distributed vector representations, or word embedding models, are mathematical representations that consist of mapping each word of a vocabulary to a dense n-dimensional vector (Bengio et al., 2003) in such a way that related words are placed close

[2] https://github.com/Petroles/Petrovec

by in the vector space (Turney and Pantel, 2010). They are based on the distributional hypothesis (Harris, 1954) which states that words appearing in a similar context tend to have similar meanings. These vector space models allow the observation of semantic similarities between words by measuring the distance between their corresponding vectors, using metrics such as the cosine similarity (Mikolov et al., 2013). This capability of encoding semantic properties has proven to be very efficient, as it can be used as features to compose the first data processing layer in deep neural network architectures, helping to achieve impressive results in several NLP tasks (Schnabel et al., 2015; Goldberg, 2016; Lai et al., 2016; Hartmann et al., 2017; Camacho-Collados and Pilehvar, 2018; Young et al., 2018; Tshitoyan et al., 2019).

The seminal research on learning distributed vector representations are by Rumelhart et al. (Rumelhart et al., 1986). Bengio et al. (Bengio et al., 2003) presented a neural network-based architecture for a language model using word embeddings. Collobert and Weston (Collobert and Weston, 2008) demonstrated the usage of pre-trained vectors applied to NLP tasks. More recently, Mikolov et al. (2013a, 2013b) introduced Word2vec, a computationally efficient neural network approach that enabled scaling the training process to large datasets. Word2vec gained popularity and helped to establish word embeddings as a fundamental unit in NLP algorithms. Other important word embedding models have followed, such as Global Vectors (Pennington et al., 2014) and FastText (Bojanowski et al., 2017). Finally, a new generation of contextual embeddings has achieved state-of-the-art results in several downstream applications, such as ELMo (Peters et al., 2018), ULMFit (Howard and Ruder, 2018), and BERT (Devlin et al., 2019). Contextualized models, however, significantly increase the computational requirements for training and inference compared to non-contextual embeddings (Polignano et al., 2020; Arora et al., 2020). In real-world industrial scenarios, where computational resources are limited and huge volumes of data are required to be processed in short response times, latency must also be considered in addition to accuracy. Polignano et al. (Polignano et al., 2020) encourages further analysis of the trade-off between accuracy and latency, to assess whether the improvements in accuracy justify the high increase of computational power required for running contextual models. Conversely, non-contextual models, such as `PetroVec`, are lighter and faster to train and run, presenting a great alternative at a fraction of the computational cost when compared to contextual models. Arora et al. (Arora et al., 2020) conducts a comprehensive study on comparing contextual and non-contextual embeddings, regarding their accuracy gains and increase in computational cost. They report that, in many production tasks using industry-scale data, classic pre-trained embeddings have highly competitive results compared to contextual embeddings, while the latter perform better on tasks involving complex text structures, ambiguity, and unseen vocabulary. Nevertheless, classic pre-trained embeddings have been widely used both in academia and industry, obtaining impressive results reported for many downstream applications (Ittoo et al., 2016; Blinston and Blondelle, 2017; Young et al., 2018; Nooralahzadeh et al., 2018; Cordeiro et al., 2019; Tshitoyan et al., 2019; Khabiri et al., 2019; Padarian and Fuentes, 2019).

2.2. Domain-specific O&G vocabulary

Despite some few public word embedding models already being available for Portuguese (Hartmann et al., 2017; Bojanowski et al., 2017), the O&G domain requires specialized models, capable of capturing all the semantic properties of its highly technical vocabulary. Some terms may assume a very specific meaning within the O&G domain when compared to a standard context understanding, such as '*árvore de natal*' ('*christmas tree*': a set of valves connected to the top of a well), '*falha*' ('*fault*': a break or planar surface in brittle rock across which there is observable displacement), or equipment names such as 'BOP' (acronym for blowout preventer, a large safety valve at the top of a well that may be closed if the drilling crew loses control). However, no such models exist for this domain in Portuguese.

3. Related work

Since the popularization of applying word embedding models in NLP applications, especially after several promising results in deep learning algorithms (Young et al., 2018), there has been an effort to provide good quality pre-trained representations for general purposes. Transfer learning techniques are commonly applied to reuse models originally trained on a general-domain corpus, feeding domain-specific algorithms with those pre-trained embeddings to perform a specific task (Ruder et al., 2019).

However, despite the success of generic embeddings, several studies show that training specialized embedding models from a domain-specific corpus can significantly increase the quality of their semantic representation and, hence, the accuracy of applications on downstream tasks (Lai et al., 2016; Pakhomov et al., 2016; Nooralahzadeh et al., 2018; Wang et al., 2018a; Alsentzer et al., 2019; Tshitoyan et al., 2019). Developing and evaluating domain-specific word embedding models is a major research field in NLP. Lai et al. (Lai et al., 2016) reported a study on the best practices for generating good-quality locally-trained word embeddings. The authors highlighted that "corpus domain is more important than corpus size" and recommended "choosing a corpus in a suitable domain for the desired task".

Domain-specific word embeddings have been widely used in research and industrial applications. Khabiri et al. (Khabiri et al., 2019) applied them to perform log classification and Mishra et al. (Mishra and Sharma, 2019) used them to effectively identify ambiguities in software requirements. Tshitoyan et al. (Tshitoyan et al., 2019) reported impressive results on using word embedding induced from materials science literature, capable of predicting potential discoveries of new thermoelectric materials years in advance, stating that "the quality and domain-specificity of the corpus determine the utility of the embeddings for domain-specific tasks".

The biomedical domain, due to its large availability of corpora and evaluation datasets, is one of the most active NLP research area (Jiang et al., 2015; Wang et al., 2018a; Alsentzer et al., 2019; Lee et al., 2020). Kalyan et al. (Kalyan and Sangeetha, 2020) presented a comprehensive survey on the applications of word embedding for clinical NLP.

Concerning the O&G domain, only recently few studies focused on providing specialized models, mainly because of the scarcity of publicly available corpora. Nooralahzadeh et al. (Nooralahzadeh et al., 2018) presented a study on generating and evaluating embedding models for O&G in English. They performed intrinsic evaluation considering a gold standard built upon the Schlumberger Oilfield Glossary, measuring accuracy, precision, and recall for semantically related terms. Their results outperformed a standard general-domain model on comparative evaluation, and the study highlighted that constructing domain-specific models is beneficial even considering the limited availability of specialized corpora. Padarian and Fuentes (Padarian and Fuentes, 2019) presented English word embedding models for geosciences, as well as a test suit for intrinsic evaluation. Their specialized models were compared to a general-domain public model, obtaining significant improvements on tasks such as semantic analogies and categorization.

All previously mentioned studies reiterate, for O&G applications, a domain-specific model is worthwhile even when the corpus is

Table 1
Acquired corpora for the O&G domain in Portuguese

Corpus domain	Source	Description	Sentences	Tokens
Specific	Petrobras[a]	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[b]	Bulletins and technical reports	132,955	2,136,465
		Theses and dissertations on the O&G domain	279,196	3,629,999
	IBP[c]	Proc. of the Rio O&G Conf.	89,116	1,287,223
	IBICT-BDTD [d]	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

[a] http://publicacoes.petrobras.com.br.
[b] http://www.anp.gov.br.
[c] Brazilian Petroleum, Gas and Biofuels Institute.
[d] Brazilian Institute of Information in Science and Technology - Brazilian Digital Library of Theses and Dissertations: http://bdtd.ibict.br/.

considerably smaller than a general-domain dataset. Specifically for Portuguese, we did some preliminary work on this topic and generated the only known set of Portuguese word embedding models for the O&G domain (Gomes et al., 2018). In this work, we significantly expand the corpus over which the embeddings are trained, we create a gold standard and a methodology for intrinsic and extrinsic evaluation, exploring a comparative analysis with a baseline model trained on a general-domain corpora.

4. Corpora and language models

Considering the lack of reference corpora, we first gathered a large collection of public documents in the O&G domain in Portuguese. The collection includes scientific and technical publications retrieved from major universities and leading institutions in this field, such as Petrobras (a Brazilian multinational corporation in the petroleum industry) and the Brazilian National Agency of Petroleum, Natural Gas and Biofuels (ANP) (a federal government agency responsible for the regulation of the petroleum sector). Additionally, we included a general-domain corpus in Portuguese assembled by Hartmann et al. (Hartmann et al., 2017), to analyze the quality of hybrid models (*i.e.*, containing both domain-specific and general-domain). Regarding this generic corpus, we obtained only a portion of approximately 50% of its original content, corresponding to the public part provided by the authors. In order to avoid any incorrect linguistic patterns introduced by translation processes, our corpora are composed of texts originally written in Portuguese by the aforementioned reference institutions, which natively operate in this language in the O&G domain. Finally, we compiled a representative corpus composed of public scientific articles, theses, dissertations, glossaries, periodicals, and technical reports, comprising more than 85 million tokens. Table 1 shows detailed information about the corpora, their corresponding sources, and the total tokens and sentences count after preprocessing operations. To the best of our knowledge, the domain-specific corpus is the largest body of text ever reported for the O&G domain for Portuguese.

4.1. Corpora preprocessing

The corpus was preprocessed for extracting, cleaning, and preparing the data to be in a suitable format to train the machine learning algorithms. Since most of the documents obtained during acquisition step were in PDF and MS Word formats, these files were then converted to plain text using the Java library Apache Tika (Mattmann and Zitting, 2012). After conversion, all text documents were processed for lexical normalization and noise removal, considering the best practices for Portuguese language described in Hartmann et al.(Hartmann et al., 2017) and Rodrigues et al. (Rodrigues et al., 2016). All characters were lowercased and diacritics, punctuation and special characters were removed. Numerical tokens were replaced by a <NUMBER> tag, and multiple consecutive occurrences of the same tag were replaced by a single one. Common words (*i.e.*, stopwords) were also discarded using NLTK Python library (Loper and Bird, 2002). Finally, the clean texts were tokenized into sentences and lines with fewer than three words were eliminated.

4.2. Training the word embedding models

We trained the `PetroVec` models on Word2vec (Mikolov et al., 2013) and FastText (Bojanowski et al., 2017), using the implementation from the Gensim Python library (Řehůřek and Sojka, 2010). Since our main goal is to provide a set of domain-specific pre-trained embeddings, rather than presenting technical novelty about the training procedures, we set the hyperparameters to the values recommended in previous studies (Mikolov et al., 2013; Lai et al., 2016; Hartmann et al., 2017; Nooralahzadeh et al., 2018), *i.e.*, window size = 5, skip-gram method, and a vector size of 100 dimensions, which may provide a good balance between model quality and performance. We also discarded all words that occur fewer than ten times.

To enable a comparative evaluation on the effects of different corpora on the quality of the resulting embeddings, we generated different versions of the models considering two corpora compositions, identified in Table 1 as *specific*and *hybrid*. The model trained from the specific corpus will be referred to as `PetroVec-O&G`, and the model trained from the hybrid corpus will be referred to as `PetroVec-hybrid`. For each the two corpora compositions, two embeddings algorithms were used: Word2vec and FastText. Additionally, we downloaded a public Portuguese general-domain model built by Hartmann et al. (Hartmann et al., 2017) to serve as a baseline to our evaluations, available at the word embeddings repository of the Interinstitutional Center for Computational Linguistics (NILC)[3] . This model was trained on a large general-domain

[3] http://nilc.icmc.usp.br/embeddings

Table 2
Corpora composition for each model

Model	Description	Vocabulary size	Corpus size
`baseline-O&G`	Baseline O&G model trained on a preliminary version of the domain-specific corpus (Gomes et al., 2018)	113,934	10,109,732
`PetroVec-O&G`	Model trained on the domain-specific O&G corpus	161,842	85,725,834
`PetroVec-hybrid`	Model trained on both domain-specific and generic corpora	440,692	365,295,169
`skipgram-NILC`	Pre-trained model in Portuguese from general-domain corpora (Hartmann et al., 2017)	929,606	1,395,926,282

corpus of about 1.4 billion tokens and it is henceforth referred to as `skipgram-NILC`. Finally, to enable further comparative evaluations, we also used a pre-trained O&G model in Portuguese from our preliminary work (Gomes et al., 2018), referred to as `baseline-O&G`, trained on a smaller version of the specific corpus. Table 2 shows detailed information about each model, including their corresponding corpus and vocabulary size.

5. Intrinsic evaluation

The intrinsic evaluation aims to assign a metric of how well the embeddings can encode the semantic and syntactic properties of the text. The process consists of using the models to rate the similarity of pairs of words and compare them to the human perception of similarity (Baroni et al., 2014; Schnabel et al., 2015; Gladkova and Drozd, 2016). After a thorough search, we found no evaluation datasets in the O&G domain for Portuguese. Thus, in order to create a dataset for intrinsic evaluation of word embeddings in the O&G domain, we followed an approach based on semantic relatedness (Zhang et al., 2013). The methodology is described in the next section, followed by a discussion of the results.

5.1. Dataset development methodology

The intrinsic evaluation task consists of calculating the correlation between the similarity score assigned by human subjects and the cosine similarity calculated between terms by the models being tested (Schnabel et al., 2015; Camacho-Collados and Pilehvar, 2018). To enable such a calculation, an evaluation dataset (*i.e.*, a gold standard) containing terms that are relevant to the O&G domain is needed.

Our source was a translated version of the Petroleum Abstracts' Exploration & Production Thesaurus[4] . The translation was performed manually by domain specialists from Petrobras. A set of 1,500 unique word pairs was selected from this thesaurus. In order to ensure that the evaluation dataset did not contain only unrelated word pairs, the choices were curated in such a way that resulted in a balance between highly related words, partly related words, and unrelated words. The three types of relations on the thesaurus (*broader*, *narrower*, and *related*) were used to guide this selection. Highly related words amounted to 25% of the selected pairs and were randomly taken from the *narrower* and *broader* relationships. Half of the randomly selected pairs were *related*, and the remaining 25% of the pairs were composed of unrelated words (*i.e.*, words that were not connected by any of the relations from the thesaurus).

The annotators were 25 undergraduate and graduate students, one PhD in geoscience, as well as ten specialists from the O&G industry. The word pairs were annotated by the human subjects using two different methods: (i) selection of the most semantically related between two pairs (used as support for the quality of judgment), and (ii) assignment of a value from 1 to 7 in a Likert scale (Likert, 1932) to reflect degree of relatedness for each pair of words (used for the resulting evaluation dataset).

Table 3
The correlation between annotators

Annotators	Spearman
1 and 2	0.68
2 and 3	0.63
1 and 3	0.68
Mean	0.66

For annotation method (i), 300 of the 1,500 pairs were selected at random. Each of the 300 pairs was individually compared to 24 other randomly chosen pairs, also within the 300 subset. These comparisons were pair-to-pair, and every time a specific pair was chosen as the most related of the two, it scored a point. Once the annotation was completed, each pair in the set had a score ranging from 0 to 24, according to how many times they were deemed the most related. Pairs with higher scores contain words that are more related. Conversely, those with lower scores, have words that are less related. Although each of the pair-to-pair comparisons were made only once, each pair was compared to other pairs 24 times, this ensures the generality of the score. In the end, 3,600 pair-to-pair comparisons were performed in order to annotate 300 word pairs. This annotation method was inspired by the works by Aguirre et al. (Agirre et al., 2009) and Bruni et al. (Bruni et al., 2014).

For annotation method (ii), all 1,500 semantic pairs were scored from 1 to 7 (where 7 represents the highest relatedness degree). Each semantic pair was annotated three times using this method, once by a geoscience graduate student, once by a PhD in geoscience, and once by a specialist from the O&G industry. The reference score to which the semantic models were compared to is the average score of the three annotations. The agreement between annotators, calculated by the Spearman correlation, can be found in Table 3.

The goal of annotating with these two methods was to ensure the adequacy of semantic relatedness annotation in the resulting evaluation dataset by analyzing the correlation between these two forms of annotation. The first one is based on judgments that are easier to make (which pair from two choices is more semantically related), but it requires annotating a greater volume of items. The second method (*i.e.*, grading relatedness from 1 to 7) is less intuitive for the annotators, but it takes fewer annotations. In order to guarantee that the two types of analysis were correlated, we performed a test on a subset of the data (300 word pairs) for which we had annotations for both methods. We found a Spearman's ρ of 0.86 between these methods, which indicates that the annotation based on the Likert scale was fair and fit to be used as the evaluation dataset for our intrinsic evaluation. Annotation instructions, which included examples of how annotators should interpret the Likert scale, and resulting evaluation dataset are provided in our public repository [5].

Besides the correlation with human annotation, we also measured the *coverage*, *i.e.*, the fraction of the word pairs that are present in the model. Coverage is important to measure how much of the vocabulary is represented by the model.

[4] https://www.pa.utulsa.edu/products/tulsadatabase/thesaurus

[5] https://github.com/Petroles/Petrovec

Table 4
Results for the intrinsic evaluation (semantic similarity test). Best results for each metric are in bold

Model	Algorithm	Spearman (ρ)	Coverage	Harmonic Mean
baseline-O&G	FastText	0.43	0.89	0.58
	Word2vec	0.51	**1.00**	0.68
skipgram-NILC	FastText	0.57	0.94	0.71
	Word2vec	0.48	0.94	0.64
PetroVec-O&G	FastText	0.56	**1.00**	0.72
	Word2vec	0.61	**1.00**	0.76
PetroVec-hybrid	FastText	0.61	**1.00**	0.76
	Word2vec	**0.65**	**1.00**	**0.79**

Finally, in order to rank the results of the four models (baseline-O&G, skipgram-NILC, PetroVec-O&G, and PetroVec-hybrid), it is important to consider correlation and coverage simultaneously since the best model should capture semantic similarity and also include the necessary terminology. Therefore, in order to aggregate these two notions into a single metric, we calculated the *harmonic mean* between correlation and coverage. The rationale was to simulate the behavior of the widely used F1.

5.2. Results

To assess how well the models have learned domain semantics, they were used to calculate a similarity measure between each of the 1,500 previously mentioned word pairs, and this list of model-based similarities was compared for correlation with our Likert-annotated evaluation dataset. Table 4 presents the correlation, coverage, and the harmonic mean obtained for each of the trained models. Our results show that general-domain corpora are inferior to domain-specific corpora when it comes to coverage, as expected. It also confirms that models trained on larger corpora have better results. This can be seen by the superiority of PetroVec-O&G in comparison to baseline-O&G. It is, however, interesting to see that for domain-specific models, FastText-based models performed slightly worse than Word2vec-based models, the opposite of what can be seen in the NER results, examined in the next section. The reason behind why the general-domain model did not follow this trend might be explained by the fact that skipgram-NILC's FastText model is distributed in a text format that loses some characteristics of the FastText architecture, making it more similar to the Word2Vec model when put to use.

To assess whether the differences found in the dependent correlation coefficients were significant, following (Faruqui et al., 2016; Shalaby and Zadrozny, 2017), we calculated the Steiger's Z test (Steiger, 1980). The results showed that PetroVec-O&G using Word2vec is significantly better than baseline-O&G and skipgram-NILC (using *alpha* = 0.05) using both embedding algorithms. The best overall model was PetroVec-hybrid using Word2vec as it significantly outperformed all other models.

Faruqui et al. (Faruqui et al., 2016) points out that the use of word similarity tasks as a proxy for the intrinsic evaluation of word embeddings presents several limitations and recommend that word vectors should be evaluated on downstream tasks. Such an evaluation is the focus of the next section.

6. Extrinsic evaluation

Extrinsic evaluations measure the contribution of a word embedding model when used as input for specific NLP tasks (Turian et al., 2010; Schnabel et al., 2015), such as automatic text classification, named entity recognition (NER) or part-of-speech tagging. In this work we perform an evaluation for the task of NER in Geoscience related literature. In the next sections, we report on the methodology and the results.

6.1. Evaluation methodology

For the extrinsic evaluation task, NER in geoscience, we adopted a revised version of the GeoCorpus, first presented in (Amaral, 2017). GeoCorpus is an annotated named entity recognition (NER) resource containing the identification of relevant entities for the sedimentary basin domain. The original GeoCorpus was revised by us, being improved in several ways. For instance, the large and heterogeneous "other" category was replaced with several new, specific categories such as "Magmatic Rock". This revision was done with the assistance of a PhD in geosciences. As the sedimentary basin domain is a subdomain of O&G, we judged that our domain embeddings should be able to improve the efficiency of this task. All changes are detailed on our Github page [6].

The complete revised corpus is composed of 5,275 sentences, with 8,933 named entities divided into 30 categories. To ensure more lexical diversity, we selected the ten categories with the largest variety and highest number of named entities. A table showing the full spread of categories can be found in the Section 1.1 of the supplementary material. We selected sentences that contained at least one instance of ten most frequent categories, amounting to of 2,978 sentences, with 6,578 named entities. Details are presented in Table 6, wherein the individual class results for the best performing model are also given.

As with the intrinsic evaluation, we performed comparative analyses of the models induced from different corpora to identify the best performing models. A state-of-the-art NER neural network architecture, provided by Santos et al.Sa (Santos et al., 2019), was fed with the alternative embeddings, and measured the difference in performance, with the skipgram-NILC model serving as our baseline. This evaluation was performed via cross-validation. For this purpose, the corpus was shuffled into 10 randomized configurations of training set (roughly 70% of the corpus), testing set (roughly 20% of the corpus), and validation set (roughly 10% of the corpus), preserving a fair distribution of the named entity instances. Each configuration has different sections of the corpus randomly assigned to training, testing and validation in order to ensure as fair a test as possible.

6.2. Results

Table 5 presents the mean results for a $k = 10$ cross-validation for each of the eight tested models. The best results were obtained by PetroVec-O&G and PetroVec-hybrid (both using FastText). This superiority was due to the higher recall, which is a consequence of the wider domain coverage.

Regarding the training corpora, a comparison between skipgram-NILC and PetroVec-O&G shows that in-domain corpora bring gains in terms of vocabulary coverage, which enabled us to achieve better results with a training corpus an order of magnitude smaller (*i.e.*, skipgram-NILC is roughly 15 times larger than PetroVec-O&G). Nevertheless, size is important – as can be seen comparing PetroVec-O&G and baseline-O&G. skipgram-NILC's results, especially its precision scores, may also be attributed to training corpus size, as even without training in specialized texts it managed to learn enough about general Portuguese semantics to predict correct labeling in over 80% of cases. Adding general-

[6] https://github.com/Petroles/Petrovec

Table 5
Mean results for named entity recognition under different configurations. The best results by metric are in bold.

Model	Algorithm	Precision	Recall	F1
baseline-O&G	FastText	0.81	0.75	0.78
	Word2vec	0.78	0.69	0.74
skipgram-NILC	FastText	**0.83**	0.82	0.83
	Word2vec	0.80	0.80	0.80
PetroVec-O&G	FastText	**0.83**	**0.89**	**0.86**
	Word2vec	0.82	0.81	0.81
PetroVec-hybrid	FastText	**0.83**	**0.89**	**0.86**
	Word2vec	0.82	0.81	0.82

Table 6
Categories, number of instances, and NER results for the PetroVec-O&G model.

Named Entity Category	Instances	Precision	Recall	F1
Rocks				
Siliciclastic Sedimentary Rock	1098	0.85	0.88	0.86
Magmatic Rock	580	0.86	0.95	0.91
Metamorphic Rock	377	0.83	0.86	0.85
Carbonate Sedimentary Rock	355	0.79	0.87	0.83
Time				
Age	796	0.76	0.97	0.85
Period	712	0.92	0.92	0.92
Epoch	686	0.90	0.90	0.90
Stratigraphic Elements				
Stratigraphic Unit	763	0.81	0.86	0.83
Sedimentary Basin	551	0.86	0.88	0.86
Places				
Basin's Geological Context	660	0.75	0.83	0.79
All	6578	0.83	0.89	0.86

domain content to the training corpora of PetroVec-hybrid, however, did not significantly impact its evaluation metrics.

Comparing Word2vec and FastText, we found that FastText aided in enhancing the recall results for PetroVec-O&G and PetroVec-hybrid models by eight percentage points over their Word2vec counterparts. The Word2vec architecture did not manage to take advantage of the more specialized vocabulary present in the smaller domain corpus. The reason may be related to the less flexible nature of the Word2Vec architecture, as it works on the word level. Finally, the results seem to show that FastText's subword information can make better use of the properties of a domain-specific corpus.

In order to show the statistical significance of these values through t-tests, we first performed the Kolmogorov–Smirnov normality test on each series of results. These confirm that our data does not significantly differ from that which in normally distributed (with divergence values ranging from 0.13 to 0.29 and p-values ranging from 0.31 to 0.99). The subsequent t-test showed that the gain in recall of PetroVec-O&G compared to skipgram-NILC is statistically significant (p-value = 0.001).

Comparing the results for F1, we found that all differences in scores that are greater than 0.01 were considered significant using $\alpha = 0.05$. In this sense, no differences were found between the following pairs of models: ⟨skipgram-NILC FastText, PetroVec-hybrid Word2vec⟩, ⟨PetroVec-hybrid Word2vec, PetroVec-O&G Word2vec⟩, and ⟨PetroVec-O&G FastText, PetroVec-hybrid FastText⟩.

Table 6 presents the mean results by category for one of the models that achieved the best NER results, *i.e.*, PetroVec-O&G FastText. This model achieved mostly uniformly good results for precision and recall, with a single significant outlier: the recall result for the age category is 0.97, almost 0.1 above the average, and its precision result is 0.76, almost 0.1 below the average. Analyzing the named entities in this category, we found that they commonly end in one of the following tetragrams: "*iano*"/"*iana*", an example being "Artinskiano" (Artinskian); and "*eano*"/"*eana*", an example being "Zancleano" (Zanclean). We believe that FastText's feature of using subword information may have contributed to these results, allowing the model to recognize this category particularly well, but also causing it to mislabel unrelated words possessing the tetragrams. Relatedly, given the morphology of the O&G domain vocabulary, we believe that representing words as a bag of character n-grams is still useful for correctly encoding rare technical terms on this domain. In addition, it can also be suitable for handling out-of-vocabulary words, identifying common morphemes and taking advantage of the Portuguese morphological structure and its different inflected forms (Bojanowski et al., 2017).

7. Qualitative evaluation

In addition to the aforementioned intrinsic and extrinsic evaluations, we conducted some experiments on qualitative analyses of semantic relatedness for sets of terms representing the O&G technical vocabulary. These evaluations include word analogies, semantic space coherence and categorization (Turian et al., 2010; Schnabel et al., 2015).

7.1. Word analogy

Word analogy tests aim to find a term y for a given term x so that x: y best resembles a sample relationship a: b (Schnabel et al., 2015). Mikolov et al. (2013a) demonstrated the existence of certain linguistic regularities in the embedding space, so that word embedding models can be used to resolve subtle semantic relationships between terms, such as "*France is to Paris as Germany is to ?*", which the model is expected to return the word "*Berlin*". This analogy can be resolved as an algebraic operation, calculating the cosine similarity from the difference of the sampled word vectors (Eq. (1)):

$$\frac{(v_b - v_a)^\top . (v_y - v_x)}{||v_b - v_a|| \, ||v_y - v_x||} \tag{1}$$

Therefore, we explored analogy operations to investigate how well the models can encode linguistic relations from the O&G technical vocabulary. First, we referred to the Dictionary of Petroleum[7] to select three categories from related subdomains: (i) geology, (ii) technical professions, and (iii) measures and instruments. For each category, we provided a word-pair as a sample and polled PetroVec for additional query words to be evaluated as analogy operations (*i.e.*, in the format a is to b as x is to y), as described in Eq. (1). For instance, for the *technical professions* category, given the example *'geologo' is to 'geologia'*, we calculated analogy operations for each query term, and obtained the first word closest to the resulting vector. Furthermore, we compared our results to the skipgram-NILC for every analogy operation, and the complete results are presented in Table 7. The models are polled for the Portuguese terms, whereas the English translations are provided for the reader's understanding, as defined by Dictionary of Petroleum. We found that our specialized model correctly retrieves the expected terms most related to the analogy operations, maintaining its semantic relation within the O&G domain. In contrast, the skipgram-NILC model obtains good results for most of the terms in the *technical professions* category, but fails on the other subdomains. These results suggest that our specialized model was

[7] http://dicionariodopetroleo.com.br/

Table 7
Examples of relationships between word pairs retrieved from the specialized and generic models, given the examples on the header, in the form *a* is to *b* as *x* is to *?*

Geology terms	Technical professions	Measures and instruments
silte - siltito	geologo - geologia	temperatura - termometro
(silt - siltstone)	*(geologist - geology)*	*(temperature - thermometer)*
	PetroVec	
folhelho - arenito	geofisico - geofisica	pressao - manometro
(shale - sandstone)	*(geophysicist - geophysics)*	*(pressure - pressure gauge)*
argila - argilito	engenheiro - engenharia	volume - bureta
(clay - mudstone)	*(engineer - engineering)*	*(volume - burette)*
calcario - dolomito	quimico - quimica	direcao - bussola
(limestone - dolomite)	*(chemist - chemistry)*	*(direction - compass)*
sal - evaporito	gestor - gestao	porosidade - porosimetro
(salt - evaporite)	*(manager - management)*	*(porosity - porosimeter)*
feldspato - hornblenda	oceanografo - oceanografia	densidade - densimetro
(feldspar - hornblende)	*(oceanographer - oceanography)*	*(density - densimeter)*
carbonato - anidrita	enfermeiro - enfermagem	corrente - multimetro
(carbonate - anhydrite)	*(nurse - nursing)*	*(current - multimeter)*
	skipgram-NILC	
folhelho - lagerstätte	geofisico - geofísica	pressao - apetite
(shale - lagerstätte)[a]	*(geophysicist - geophysics)*	*(pressure - appetite)*
argila - lacados	engenheiro - engenharia	volume - negócio
(clay - lacados)[a]	*(engineer - engineering)*	*(volume - business)*
calcario - lobenstein	quimico - química	direcao - c-leg
(limestone - lobenstein)[a]	*(chemist - chemistry)*	*(direction - c-leg)*[a]
sal - meloa	gestor - gerência	porosidade - dielétrico
(salt - meloa)[a]	*(manager - sector)*	*(porosity - dielectric)*
feldspato - feldspatos	oceanografo - oceanografia	densidade - dólar-turismo
(feldspar - feldspars)	*(oceanographer - oceanography)*	*(density - tourism dollar)*
carbonato - hidróxido	enfermeiro - odontologia	corrente - modismo
(carbonate - hydroxide)	*(nurse - odontology)*	*(current - fad)*

[a] Some terms do not have a corresponding translation within the O&G domain, as they reflect noise and incorrect return from the generic model.

Table 8
Examples of relationships between bilingual word pairs correctly retrieved by PetroVec, given the example *'reservatorio' -'reservoir'*

Portuguese-English analogies Given: *reservatorio - reservoir*	
exploracao - exploration	producao - production
perfuracao - drilling	sismica - seismic
campo - field	bacia - basin
plataforma - platform	oleo - oil
hidrocarboneto - hydrocarbon	combustivel - fuel
duto - pipe	rocha - rock
falha - fault	poco - wells
porosidade - porosity	permeabilidade - permeability
viscosidade - viscosity	pressao - pressure
arenito - sandstone	sal - salt
sedimento - sediment	brasil - brazil

able to encode some syntactic properties, as observed in terms containing the same suffix (*e.g.*, the suffix '*metro*' as in 'manometro', 'densimetro' and 'porosimetro' from the category *measures and instruments*), and also more advanced semantic relations that do not share any stems or suffixes (as in 'bureta' and 'bussola' from the same category).

Interestingly, we found that PetroVec was able to correlate some technical terms in Portuguese to their corresponding translation in English using analogy operations. This has happened even with the selection of corpora focusing mainly in Portuguese, without any explicitly designed parallel bilingual content. Table 8 describes some advanced analogies from bilingual terms selected from Dictionary of Petroleum and correctly identified by the model, given only the example '*reservatorio' is to 'reservoir*'. Additionally, following Mikolov et al. (2013a), Fig. 1 shows a two-dimensional PCA projection of these analogies. To confirm that the cross-lingual similarity was not due to a corpus artifact, we computed the co-occurrences of the cross-lingual word pairs in Table 8 within a window of 5 words (as this was the parameter used in embedding generation). The results showed that co-occurrence is quite rare, with the most frequent pair appearing in the same window only 0.44% of the times. In addition, following Linzen (Linzen, 2016), we performed additional analyses for the word analogies, assessing whether the quality of the method is affected by reversing the direction of each analogy operation. We found that, despite a minor decrease in quality, the reversed analogies are mostly consistent, having the expected term correctly listed as one of the top-5 closest results in the vast majority of the cases (Rogers et al., 2017; Newman-Griffis et al., 2017). The complete results are presented in Section 1.2 of the Supplementary Material.

However, we highlight that this behavior is not extensive to all possible translations, as PetroVec is not intended to be a translation tool itself. Nevertheless, we were able to successfully replicate these analogies for a large number of terms from the domain-specific vocabulary. We further investigated some additional scenarios for these translation analogies, expanding the results to list the top-3 words most related to the resulting vector, allowing for multiple correct answers (Newman-Griffis et al., 2017) and hence exploring a broader neighborhood area. Table 9 describes some examples where PetroVec fails to find the correct word as the first result, whereas in many cases the expected term is listed as one of its nearest neighbors. Also, there are cases in which the expected word is not listed, but the retrieved terms are all semantically related to the expected translation. Moreover, in some particular cases, the model produces more than one acceptable translation for the query term. Regardless, this translation feature of the semantic space can provide great generalization capabilities for those bilingual concepts, which is crucial to a variety of downstream tasks, such as semantic search, question answering, conversational platforms and machine translation.

Finally, we emphasize that word analogy evaluations have several limitations on assuming linguistic regularities that may not hold in real-word data (Rogers et al., 2017; Newman-Griffis et al., 2017; Linzen, 2016), since natural language semantics is more com-

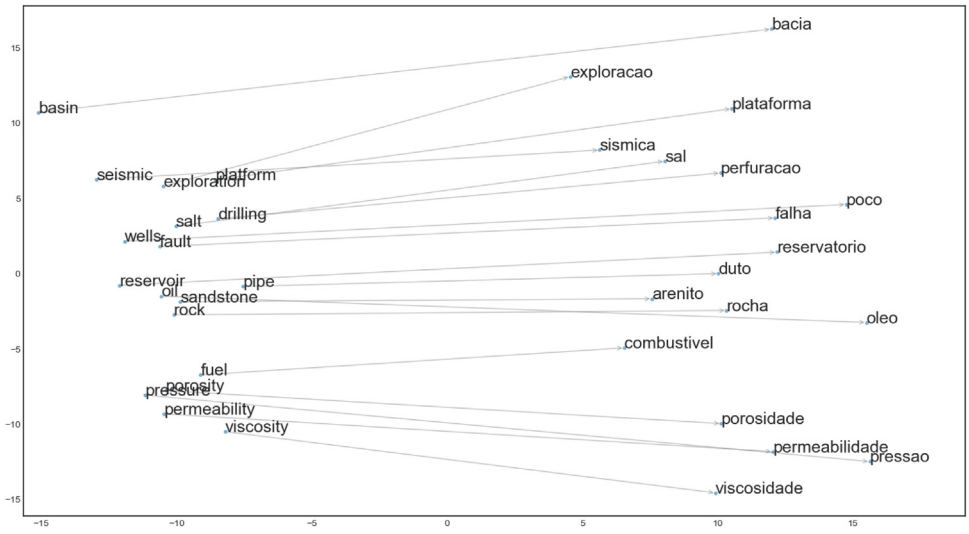

Fig. 1. Two-dimensional PCA projection of analogies that identify English-Portuguese translations.

Table 9
Expanded neighborhood tests for the resulting vectors of the analogy operation (the expected results are in bold), given the example *'reservatorio' - 'reservoir'*

Portuguese-English analogies Given: *reservatorio - reservoir*	
particula - *frsg, droplet,* ***particle***	completacao - *well,* ***completion****, drilling*
batimetria - *depth, seismic, spatial*	poros - *melim, vuggy,* ***pore***
geofisica - *seismic, geopro, sulfabras*	cimentacao - *completion,* ***cementing****,* ***cementation***
asfalto - *speedy, brasivil,* ***asphalt***	navio - *moored,* ***ship****, operations*
camada - *wiu,* ***layer****, caniada*	estimativa - ***estimation****,* proxy, ***prediction***
risco - ***risk****,* probability, attractiveness	perfil - *perl,* ***profile****,* ***logs***
planta - ***plant, facility****,* microplanta	prospeccao - *seismic, exploration, exploratory*
bomba - *centrifugal,* ***pump****, microvalve*	condutividade - *permeability, density,* ***conductivity***

Table 10
Sample technical terms for each O&G subdomain (the models are polled for the Portuguese terms, whereas the English translation is provided for the reader's understanding)

Geology	**Drilling**	**Geophysics**	**Petroleum derivatives**
anidrita *(anhydrite)*	valvula *(valve)*	eletromagnetica *(electromagnetic)*	nafta *(naphtha)*
evaporito *(evaporite)*	duto *(pipe)*	espectral *(spectral)*	querosene *(kerosene)*
calcarenito *(calcarenite)*	riser *(riser)*	amplitude *(amplitude)*	glp *(lpg)*
argila *(clay)*	pig *(pig)*	reflexao *(reflection)*	gasolina *(gasoline)*
arenito *(sandstone)*	choke *(choke)*	acustica *(acoustic)*	diesel *(diesel)*
folhelho *(shale)*	anm *(christmas tree)*	sismica *(seismic)*	combustivel *(fuel)*

plex than what can be represented as linear analogy operations. Hence, we believe that word analogies are not sufficient on their own to determine the quality of word embeddings, and should be combined with other types of analyses to provide significant evidence for a consistent evaluation.

7.2. Semantic space coherence

Considering the aforementioned lack of reference benchmarks and annotated test sets for this domain and language, we chose to perform semantic space coherence evaluation based on qualitative analysis to explore some of its geometric properties, instead of estimating quantitative metrics based on approaches such as the *intruder word* concept (Schnabel et al., 2015) or outlier detection (Camacho-Collados and Pilehvar, 2018).

Schnabel et al. (Schnabel et al., 2015) suggested that a proper semantic space should be organized in order to provide coherent neighborhoods for each word vector (Gladkova and Drozd, 2016). That is, the word embedding model should be able to create clusters of semantically similar items (Camacho-Collados and Pilehvar, 2018). Hence, we start investigating semantic space coherence by conducting a qualitative analysis of the multidimensional space in a two-dimensional projection, using the t-SNE (Maaten and Hinton, 2008) algorithm for dimensionality reduction from 100 to 2 dimensions. Following a similar approach as reported by Tshitoyan et al. (Tshitoyan et al., 2019), Wang et al. (Wang et al., 2018a), and Camacho-Collados and Pilehvar (Camacho-Collados and Pilehvar, 2018), the goal is to provide an intuitive overview of the embedding space, assessing whether small groups of words are mutually related. Fig. 2 shows the t-SNE projection of the word embedding space for the 20,000 most frequent words from our domain-specific corpus. The highlighted area focus on the word *'jurassico' (jurassic)*, which corresponds to a geological epoch. Note that the nearest neighbors are all related to the same concept, suggesting that the model was able to correctly encode the similarity of such related terms, assigning their vectors to a nearby position with notable cohesion.

Further, we performed a more detailed analysis of semantic space coherence, evaluating if the model is capable of retaining cohesion when providing groups of semantically related words. We resorted to the Schlumberger Oilfield Glossary[8] to elect four relevant O&G subdomains: (i) *geology*, (ii) *drilling*, (iii) *geophysics*, and (iv) *petroleum derivatives*. Then, for each subdomain, we selected six words from the Oilfield Glossary, referring to the Dictionary of Petroleum to obtain their correct representations in Portuguese. Table 10 lists the selected subdomains and their corresponding terms. We polled `PetroVec` for the cosine similarity between each possible pair of words, and created a similarity matrix. We pro-

[8] https://www.glossary.oilfield.slb.com

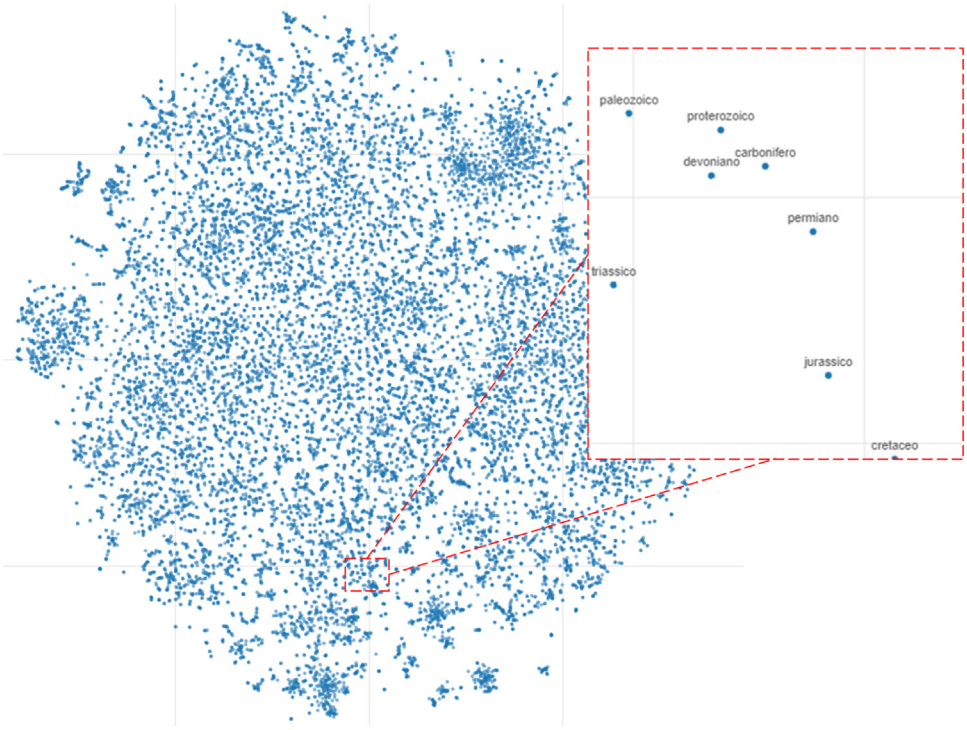

Fig. 2. Two-dimensional t-SNE projection of the semantic space for the 20,000 most common words, with highlighted neighborhood for the word '*jurassico*'.

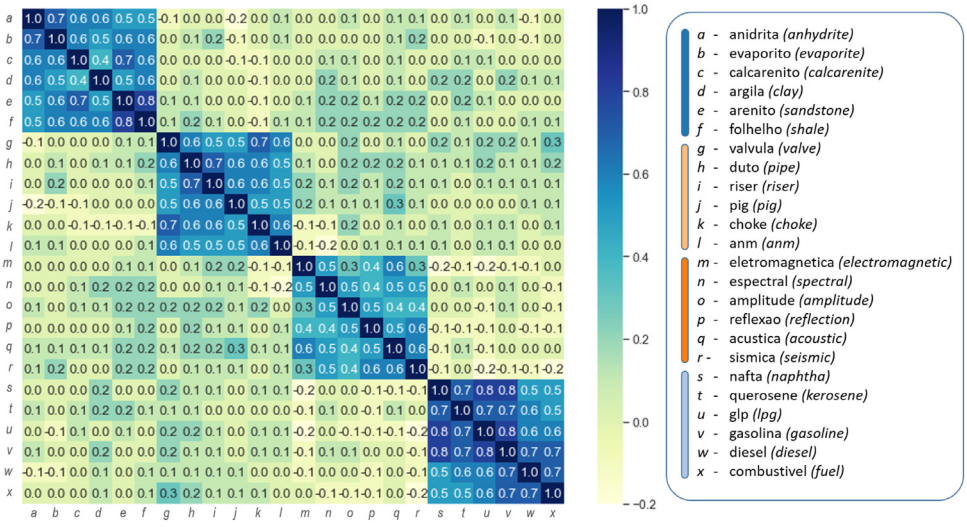

Fig. 3. Heatmap for the similarity matrix, highlighting strong clusters within the same subdomain.

posed two complementary visualizations for this similarity matrix: (i) as a heatmap chart (Fig. 3), highlighting their similarity scores depicted as color gradients; and (ii) as a chord diagram (Fig. 4), where the similarity relations are represented as the strength of the connections between elements. The heatmap provides an overall visualization of the scores between the elements, emphasizing the formation of cohesive clusters of elements within the same category. The chord diagram highlights that the interconnection between elements is predominant between terms within the same subdomain (*i.e.*, there are very few lines connecting elements from different clusters). Both charts show the formation of distinct and cohesive groups, with minor overlapping of relevant similarity scores between words from different subdomains, suggesting that the model was able to correctly encode a highly coherent semantic space.

7.3. Categorization

In addition to coherence tests, we performed categorization analyses using a clustering algorithm to investigate whether the model is able to provide coherent clusters of semantically related terms. The goal is to select a set of query words from predefined categories and their corresponding neighborhood. The result of a clustering algorithm should able to automatically distinguish the pre-established categories, assigning all the related words to their corresponding group (Schnabel et al., 2015).

For this task, we manually elected a new set of query words representing five different subdomains: (i) *carapeba* (a Brazilian production field), (ii) *combustivel* (*fuel*, a petroleum product), (iii) *paleoceno* (*paleocene*, a geological epoch), (iv) *geofisica* (*geophysics*, a subject of science relevant to the domain), and (v) *arenito* (sand-

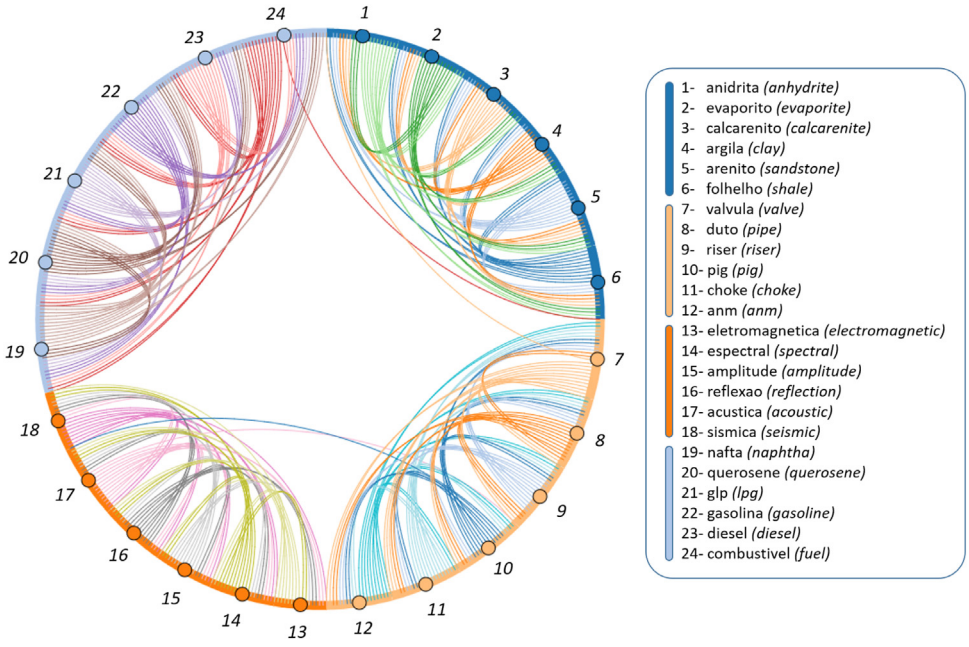

Fig. 4. Chord diagram for the similarity matrix, highlighting strong relationships between terms within the same category (relations with similarity scores greater than 0.3 are displayed, as scores below this threshold are considered irrelevant).

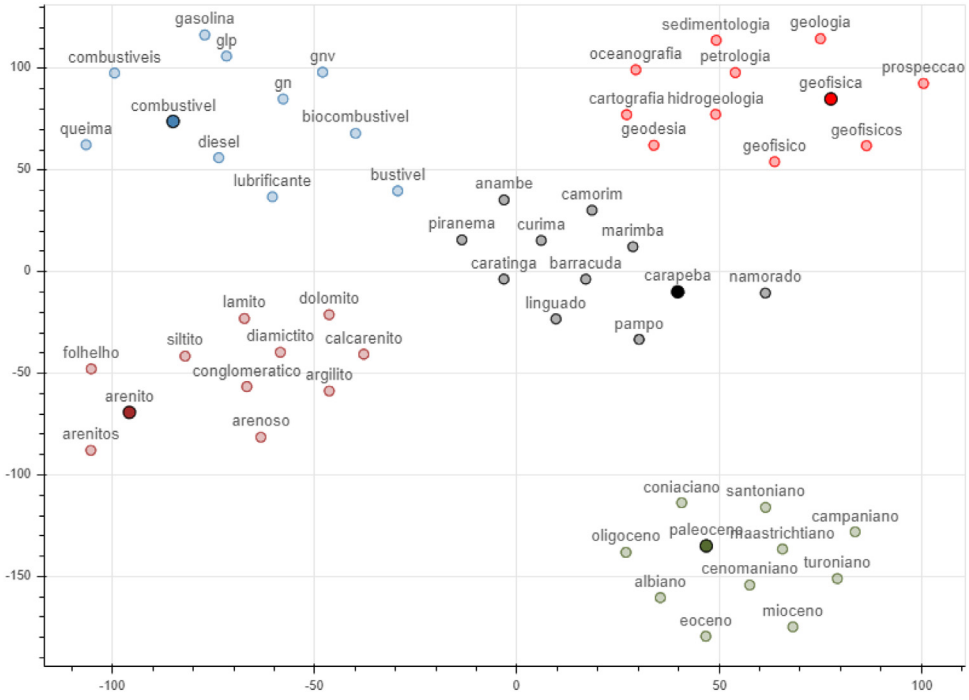

Fig. 5. t-SNE projection of k-means clustering for nearest neighbors of sample terms from different O&G subdomains.

stone, from geology domain). For each query word, we polled the model for the top-10 nearest neighbors. We stacked all the vectors and the resulting matrix was clustered using the k-means method. The results were projected into a two-dimensional space using t-SNE, as shown in Fig. 5. We can see that that k-means correctly assigned all the related words to the same group as their corresponding query term as expected, without overlapping with any other term from different categories, suggesting a significant cohesion within the clusters.

8. Conclusion

In this work, we introduced `PetroVec`, a set of domain-specific pre-trained word embedding models for the O&G industry in Portuguese. These embeddings were induced from a large collection of textual resources gathered from leading institutions in this domain. We also created an annotated test set, labeled by experts in geosciences and petroleum engineering, designed to perform intrinsic semantic metrics. The generated models

were thoroughly evaluated, both quantitatively (with intrinsic and extrinsic evaluations) and qualitatively, and compared to a general-domain public model. Further, we found that `PetroVec`, trained on domain-specific corpora, outperformed a public general-domain pre-trained embedding, suggesting that our specialized models were able to encode crucial semantic properties of the O&G technical vocabulary. All the resources developed by this work are available for public use, including the pre-trained models, the evaluation dataset, the corpus and all the scripts developed for text processing and model training.

We believe that several researchers working on the O&G domain in Portuguese may benefit from the improved semantic representations these public models can provide. Some examples of the broad range of NLP applications of interest to the O&G domain include risk prediction and safety (Birnie et al., 2019; Cai et al., 2019), information extraction (Blinston and Blondelle, 2017; Wang et al., 2018b), drilling operations improvement (Wilson, 2017; Castiñeira et al., 2018; Colombo et al., 2019; Ucherek et al., 2020), technical documents summarization (Correia Marques et al., 2019), question answering systems (Jacobs, 2019), and automatic text classification (Sanchez-Pi et al., 2014; Nooralahzadeh et al., 2018; Khabiri et al., 2019; Ribeiro et al., 2020).

In future work, we intend to improve this first approach of gold standards for intrinsic evaluation, and also expand extrinsic evaluations with more diverse and complex tasks. Additionally, we consider that these models would benefit from using domain-specific knowledge resources, such as ontologies, to improve their semantic representation for less frequent lexical items (Pilehvar and Collier, 2016; Faruqui et al., 2015). Furthermore, given the common nature of the O&G vocabulary to be expressed as multi-word expressions (MWEs), it would certainly be worth improving these models to support MWEs (Constant et al., 2017), especially considering the availability of specialized knowledge resources (Newman-Griffis et al., 2018). Finally, we intend to pre-train contextual models with BERT using the large set of textual resources gathered by this work, as well as developing other advanced NLP applications making use of these vectors, such as NER, document classification and semantic search.

Conflict of interest

The authors declare no conflict of interest.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgments

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

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.compind.2020.103347.

References

Agirre, E., Alfonseca, E., Hall, K.B., Kravalova, J., Pasca, M., Soroa, A., 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, May 31 - June 5, 2009, Boulder, Colorado, USA, 19–27.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B., Kochut, K., 2017]. Text summarization techniques: a brief survey. Int. J. Adv. Comput. Sci. Appl.

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., McDermott, M., 2019]. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp. 72–78.

Amaral, D.O.F., 2017]. Reconhecimento de Entidades Nomeadas na Área da Geologia: Bacias Sedimentares Brasileiras, Ph.D. Thesis. Pontifícia Universidade Católica do Rio Grande do Sul http://tede2.pucrs.br/tede2/handle/tede/8035.

Arora, S., May, A., Zhang, J., Ré, C., 2020]. Contextual embeddings: when are they worth it? Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2650–2663, http://dx.doi.org/10.18653/v1/2020.acl-main.236 https://www.aclweb.org/anthology/2020.acl-main.236.

Baroni, M., Dinu, G., Kruszewski, G., 2014]. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Baltimore, Maryland, pp. 238–247.

Bast, H., Buchhold, B., Haussmann, E., 2016]. Semantic search on text and knowledge bases. Found. Trends(®) Inf. Ret. 10 (1), 119–271.

Bengio, Y., Ducharme, R., Vincent, P., Janvin, C., 2003. A neural probabilistic language model. J. Mach. Learn. Res. 3, 1137–1155.

Birnie, C.E., Sampson, J., Sjaastad, E., Johansen, B., Obrestad, L.E., Larsen, R., Khamassi, A., 2019]. Improving the quality and efficiency of operational planning and risk management with ML and NLP. In: SPE Offshore Europe Conference and Exhibition. Society of Petroleum Engineers, Aberdeen, UK.

Blinston, K., Blondelle, H., 2017]. Machine learning systems open up access to large volumes of valuable information lying dormant in unstructured documents. Lead. Edge 36 (3), 257–261.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2017]. Enriching word vectors with subword information. TACL 5, 135–146.

Bruni, E., Tran, N., Baroni, M., 2014]. Multimodal distributional semantics. J. Artif. Intell. Res. 49, 1–47.

Cai, S., Palazoglu, A., Zhang, L., Hu, J., 2019]. Process alarm prediction using deep learning and word embedding methods. ISA Transactions 85, 274–283.

Camacho-Collados, J., Pilehvar, M.T., 2018]. From word to sense embeddings: a survey on vector representations of meaning. J. Artif. Intell. Res. 63, 743–788.

Castiñeira, D., Toronyi, R., Saleri, N., 2018]. Machine Learning and Natural Language Processing for Automated Analysis of Drilling and Completion Data. Society of Petroleum Engineers.

Clavijo, W., de Almeida, E., Losekann, L., Rodrigues, N., 2019]. Impacts of the review of the Brazilian local content policy on the attractiveness of oil and gas projects. J.World Energy Law Bus. 12 (5), 449–463, http://dx.doi.org/10.1093/jwelb/jwz030.

Collobert, R., Weston, J., 2008]. A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceedings of the 25th international conference on Machine learning, ICML'08. Association for Computing Machinery, Helsinki, Finland, pp. 160–167.

Colombo, D., Pedronette, D.C.G., Guilherme, I.R., Papa, J.P., Ribeiro, L.C.F., Sugi Afonso, L.C., Presotto, J.G.C., Sousa, G.J., 2019]. Discovering patterns within the drilling reports using artificial intelligence for operation monitoring. Offshore Technology Conference Brasil, Offshore Technology Conference, Rio de Janeiro, Brazil.

Constant, M., Eryigit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., Todirascu, A., 2017]. Multiword expression processing: a survey. Comput. Linguist. 43 (4), 837–892, http://dx.doi.org/10.1162/COLI_a_00302 https://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00302.

Cordeiro, F.C., Gomes, D.S.M., Gomes, F.A.M., Texeira, R.C., 2019]. Technology intelligence analysis based on document embedding techniques for oil and gas domain. Offshore Technology Conference Brasil, Offshore Technology Conference, Rio de Janeiro, Brazil.

Correia Marques, J.M., Gagliardi Cozman, F., Ferreira dos Santos, I.H., 2019]. Automatic summarization of technical documents in the oil and gas industry. 2019 8th Brazilian Conference on Intelligent Systems (BRACIS), 431–436, iSSN: 2643-6264.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019]. BERT: pre-training of deep bidirectional transformers for language understanding. In: 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). Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186.

Fares, M., Kutuzov, A., Oepen, S., Velldal, E., 2017]. Word vectors, reuse, and replicability: towards a community repository of large-text resources. In: Proceedings of the 21st Nordic Conference on Computational Linguistics. Association for Computational Linguistics, Gothenburg, Sweden, pp. 271–276.

Faruqui, M., Dodge, J., Jauhar, S.K., Dyer, C., Hovy, E., Smith, N.A., 2015]. Retrofitting word vectors to semantic lexicons. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Denver, Colorado, pp. 1606–1615, http://dx.doi.org/10.3115/v1/N15-1184 https://www.aclweb.org/anthology/N15-1184.

Faruqui, M., Tsvetkov, Y., Rastogi, P., Dyer, C., 2016]. Problems with evaluation of word embeddings using word similarity tasks. In: Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP. Association

for Computational Linguistics, Berlin, Germany, pp. 30–35, http://dx.doi.org/10.18653/v1/W16-2506 https://www.aclweb.org/anthology/W16-2506.
Gladkova, A., Drozd, A., 2016]. Intrinsic evaluations of word embeddings: what can we do better? In: Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP. Association for Computational Linguistics, Berlin, Germany, pp. 36–42.
Goldberg, Y., 2016]. A primer on neural network models for natural language processing. J. Artif. Intell. Res. 57, 345–420.
Gomes, Diogo, Cordeiro, Fabio, Evsukoff, Alexandre, 2018. Word Embeddings in Portuguese for the Specific Domain of Oil and Gas. In: Proceedings of the Rio Oil & Gas Expo and Conference 2018. Instituto Brasileiro de Petroleo e Gas (IBP), Rio de Janeiro, Brazil.
Goodfellow, I., Bengio, Y., Courville, A., 2016]. Deep Learning, Adaptive Computation and Machine Learning. The MIT Press, Cambridge, Massachusetts.
Harris, Z.S., 1954]. Distributional structure. WORD 10 (2–3), 146–162.
Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Silva, J., Aluísio, S., 2017]. Portuguese word embeddings: evaluating on word analogies and natural language tasks. Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology, Sociedade Brasileira de Computação, Uberlândia, Brazil, 122–131.
Hirschberg, J., Manning, C.D., 2015. Advances in natural language processing. Science 349 (6245), 261–266.
Howard, J., Ruder, S., 2018]. Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Melbourne, Australia, pp. 328–339.
Ittoo, A., Nguyen, L.M., van den Bosch, A., 2016]. Text analytics in industry: challenges, desiderata and trends. Comp. Ind. 78, 96–107.
Jacobs, T., 2019]. The oil and gas chat bots are coming. J. Pet. Technol. 71 (02), 34–36.
Jiang, Z., Li, L., Huang, D., Jin, Liuke, 2015. Training word embeddings for deep learning in biomedical text mining tasks. 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 625–628.
Kalyan, K.S., Sangeetha, S., 2020]. SECNLP: a survey of embeddings in clinical natural language processing. Journal of Biomedical Informatics 101, 103323.
Khabiri, E., Gifford, W.M., Vinzamuri, B., Patel, D., Mazzoleni, P., 2019]. Industry specific word embedding and its application in log classification. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM'19. Association for Computing Machinery, Beijing, China, pp. 2713–2721.
Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., Brown, D., 2019]. Text classification algorithms: a survey. Inf. 10 (4), 150, number: 4 Publisher: Multidisciplinary Digital Publishing Institute.
Lai, S., Liu, K., He, S., Zhao, J., 2016]. How to generate a good word embedding. IEEE Intell. Syst. 31 (6), 5–14.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.
Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J., 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36 (4), 1234–1240.
Likert, R., 1932]. A technique for the measurement of attitudes. Arch. Psychol. 22 (140), 55–55.
Linzen, T., 2016]. Issues in evaluating semantic spaces using word analogies. In: Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP. Association for Computational Linguistics, Berlin, Germany, pp. 13–18, http://dx.doi.org/10.18653/v1/W16-2503 https://www.aclweb.org/anthology/W16-2503.
Loper, E., Bird, S., 2002]. NLTK: the natural language toolkit. In: Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics -, Vol. 1. Association for Computational Linguistics, Philadelphia, Pennsylvania, pp. 63–70.
Lu, H., Guo, L., Azimi, M., Huang, K., 2019]. Oil and gas 4.0 era: a systematic review and outlook. Comp. Ind. 111, 68–90.
Maaten, L.v.d., Hinton, G., 2008]. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (Nov), 2579–2605.
Manning, C.D., Schutze, H., 1999]. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, Mass.
Manning, C.D., 2015]. Computational linguistics and deep learning. Comput. Linguist. 41 (4), 701–707.
Mattmann, C.A., Zitting, J.L., 2012]. Tika in Action. Manning Publications, Shelter Island, NY, oCLC: ocn731912756.
Mikolov, T., Chen, K., Corrado, G.S., Dean, J., 2013a. Efficient estimation of word representations in vector space. 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013b. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13. Curran Associates Inc, Lake Tahoe, Nevada, pp. 3111–3119.
Mishra, S., Sharma, A., 2019]. On the use of word embeddings for identifying domain specific ambiguities in requirements. 2019 IEEE 27th International Requirements Engineering Conference Workshops (REW), 234–240.
Newman-Griffis, D., Lai, A., Fosler-Lussier, E., 2017]. Insights into analogy completion from the biomedical domain. BioNLP 2017, Association for Computational Linguistics, Vancouver, Canada, 19–28, http://dx.doi.org/10.18653/v1/W17-2303 https://www.aclweb.org/anthology/W17-2303.
Newman-Griffis, D., Lai, A.M., Fosler-Lussier, E., 2018]. Jointly embedding entities and text with distant supervision. In: Proceedings of The Third Workshop on Representation Learning for NLP. Association for Computational Linguistics, Melbourne, Australia, pp. 195–206, http://dx.doi.org/10.18653/v1/W18-3026 https://www.aclweb.org/anthology/W18-3026.
Niklaus, C., Cetto, M., Freitas, A., Handschuh, S., 2018]. A survey on open information extraction. In: Proceedings of the 27th International Conference on Computational Linguistics. Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 3866–3878.
Nooralahzadeh, F., Øvrelid, L., Lønning, J.T., 2018]. Evaluation of domain-specific word embeddings using knowledge resources. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan.
Padarian, J., Fuentes, I., 2019]. Word embeddings for application in geosciences: development, evaluation, and examples of soil-related concepts. SOIL 5 (2), 177–187.
Pakhomov, S.V., Finley, G., McEwan, R., Wang, Y., Melton, G.B., 2016. Corpus domain effects on distributional semantic modeling of medical terms. Bioinformatics, btw529.
Pennington, J., Socher, R., Manning, C., 2014]. Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, pp. 1532–1543.
Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018]. Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics, New Orleans, Louisiana, pp. 2227–2237.
Pilehvar, M.T., Collier, N., 2016]. Improved semantic representation for domain-specific entities. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing. Association for Computational Linguistics, Berlin, Germany, pp. 12–16, http://dx.doi.org/10.18653/v1/W16-2902 https://www.aclweb.org/anthology/W16-2902.
Polignano, M., de Gemmis, M., Semeraro, G., 2020]. Contextualized bert sentence embeddings for author profiling: the cost of performances. In: Gervasi, O., Murgante, B., Misra, S., Garau, C., Blečić, I., Taniar, D., Apduhan, B.O., Rocha, A.M.A.C., Tarantino, E., Torre, C.M., Karaca, Y. (Eds.), Computational Science and Its Applications - ICCSA 2020. Springer International Publishing, Cham, pp. 135–149.
Řehůřek, R., Sojka, P., 2010]. Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. ELRA, Valletta, Malta, pp. 45–50.
Ribeiro, L.C.F., Afonso, L.C.S., Colombo, D., Guilherme, I.R., Papa, J.P., 2020]. Evolving neural conditional random fields for drilling report classification. J. Petrol. Sci. Eng. 187, 106846.
Rodrigues, J., Branco, A., Neale, S., Silva, J., 2016]. LX-DSemVectors: distributional semantics models for portuguese. In: Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (Eds.), Computational Processing of the Portuguese Language, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 259–270.
Rogers, A., Drozd, A., Li, B., 2017]. The (too many) problems of analogical reasoning with word vectors. In: Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017). Association for Computational Linguistics, Vancouver, Canada, pp. 135–148, http://dx.doi.org/10.18653/v1/S17-1017 https://www.aclweb.org/anthology/S17-1017.
Ruder, S., Vulić, I., Søgaard, A., 2019]. A survey of cross-lingual word embedding models. J. Artif. Intell. Res. 65, 569–631.
Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986]. Learning representations by back-propagating errors. Nature 323 (6088), 533–536.
Sanchez-Pi, N., Martí, L., Garcia, A.C.B., 2014. Text classification techniques in oil industry applications. In: Herrero, Á., Baruque, B., Klett, F., Abraham, A., Snášel, V., de Carvalho, A.C., Bringas, P.G., Zelinka, I., Quintián, H., Corchado, E. (Eds.), International Joint Conference SOCO'13-CISIS'13-ICEUTE'13. Springer International Publishing, Cham, pp. 211–220.
Santos, J., Consoli, B.S., dos Santos, C.N., Terra, J., Collovini, S., Vieira, R., 2019. Assessing the impact of contextual embeddings for portuguese named entity recognition. IEEE, pp. 437–442, http://dx.doi.org/10.1109/BRACIS.2019.00083.
Schnabel, T., Labutov, I., Mimno, D., Joachims, T., 2015. Evaluation methods for unsupervised word embeddings. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Lisbon, Portugal, pp. 298–307.
Shalaby, W., Zadrozny, W., 2017]. Mined semantic analysis: a new concept space model for semantic representation of textual data. 2017 IEEE International Conference on Big Data (Big Data), 2122–2131.
Steiger, J.H., 1980. Tests for comparing elements of a correlation matrix. Psychological bulletin 87 (2), 245.
Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K.A., Ceder, G., Jain, A., 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature 571 (7763), 95–98.
Turian, J., Ratinov, L.-A., Bengio, Y., 2010]. Word representations: a simple and general method for semi-supervised learning. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Uppsala, Sweden, pp. 384–394.
Turney, P.D., Pantel, P., 2010]. From frequency to meaning: vector space models of semantics. J. Artif. Intell. Res. 37, 141–188.
Ucherek, J., Lawal, T., Prinz, M., Li, L., Ashok, P., van Oort, E., Gobert, T., Mejia, J., 2020. Auto-Suggestive Real-Time Classification of Driller Memos into Activity Codes Using Natural Language Processing. Society of Petroleum Engineers.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17. Curran Associates Inc, Long Beach, California, USA, pp. 6000–6010.

Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., Kingsbury, P., Liu, H., 2018a. A comparison of word embeddings for the biomedical natural language processing. Journal of Biomedical Informatics 87, 12–20.

Wang, C., Ma, X., Chen, J., Chen, J., 2018b]. Information extraction and knowledge graph construction from geoscience literature. Comput. Geosci. 112, 112–120.

Wilson, A., 2017]. Natural-language-processing Techniques for Oil and Gas Drilling Data. J. Pet. Technol. 69 (10), 96–97.

Yadav, V., Bethard, S., 2018]. A survey on recent advances in named entity recognition from deep learning models. In: Proceedings of the 27th International Conference on Computational Linguistics. Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 2145–2158.

Young, T., Hazarika, D., Poria, S., Cambria, E., 2018]. Recent trends in deep learning based natural language processing [review article]. IEEE Comput. Intell. Mag. 13 (3), 55–75.

Zhang, Z., Gentile, A.L., Ciravegna, F., 2013]. Recent advances in methods of lexical semantic relatedness–a survey. Nat. Lang. Eng. 19 (4), 411–479.

Zhang, L., Wang, S., Liu, B., 2018]. Deep learning for sentiment analysis: a survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 8 (4), 1253, http://dx.doi.org/10.1002/widm.1253.