Applied Computing and Geosciences 14 (2022) 100084

Contents lists available at ScienceDirect

Applied Computing and Geosciences

journal homepage: www.sciencedirect.com/journal/applied-computing-and-geosciences

Geoscience language models and their intrinsic evaluation

Christopher J.M. Lawley [a,*], Stefania Raimondo [b], Tianyi Chen [b], Lindsay Brin [b], Anton Zakharov [b], Daniel Kur [b], Jenny Hui [b], Glen Newton [a], Sari L. Burgoyne [a], Geneviève Marquis [a]

[a] Natural Resources Canada, Geological Survey of Canada, 601 Booth Street, Ottawa, Ontario, K1A 0E8, Canada
[b] ServiceNow, 161 Bay Street, Suite 13000, Toronto, Ontario, M5J 2S1, Canada

ARTICLE INFO

Keywords:
Word embedding
Language models
Machine learning
Artificial intelligence
BERT
GloVe

ABSTRACT

Geoscientists use observations and descriptions of the rock record to study the origins and history of our planet, which has resulted in a vast volume of scientific literature. Recent progress in natural language processing (NLP) has the potential to parse through and extract knowledge from unstructured text, but there has, so far, been only limited work on the concepts and vocabularies that are specific to geoscience. Herein we harvest and process public geoscientific reports (i.e., Canadian federal and provincial geological survey publications databases) and a subset of open access and peer-reviewed publications to train new, geoscience-specific language models to address that knowledge gap. Language model performance is validated using a series of new geoscience-specific NLP tasks (i.e., analogies, clustering, relatedness, and nearest neighbour analysis) that were developed as part of the current study. The raw and processed national geological survey corpora, language models, and evaluation criteria are all made public for the first time. We demonstrate that non-contextual (i.e., Global Vectors for Word Representation, GloVe) and contextual (i.e., Bidirectional Encoder Representations from Transformers, BERT) language models updated using the geoscientific corpora outperform the generic versions of these models for each of the evaluation criteria. Principal component analysis further demonstrates that word embeddings trained on geoscientific text capture meaningful semantic relationships, including rock classifications, mineral properties and compositions, and the geochemical behaviour of elements. Semantic relationships that emerge from the vector space have the potential to unlock latent knowledge within unstructured text, and perhaps more importantly, also highlight the potential for other downstream geoscience-focused NLP tasks (e.g., keyword prediction, document similarity, recommender systems, rock and mineral classification).

1. Introduction

Natural language processing (NLP) is the branch of artificial intelligence that is developing the predictive text tools that billions of people use everyday, including search, machine translation, sentiment analysis, and voice assistants (Bengio et al., 2000; Chowdhary, 2020; Hirschberg and Manning, 2015). The vast majority of these predictive text tools are based on statistical language modelling, which captures the probability distribution of words as numerical vectors (Hirschberg and Manning, 2015). These vectoral representations of words, called word embeddings, are often trained using self-supervised machine learning methods on large and unlabelled text datasets like Wikipedia. Word embeddings can be constructed from the co-occurrence frequencies of words in these training corpora, based on the basic assumption that words occurring together, or in similar contexts, tend to be more closely related (Mikolov et al., 2013a, 2013b; Pennington et al., 2014). Countries and their capital cities represent the canonical example of this proximity relationship (Mikolov et al., 2013a; 2013b). Static word embeddings like One Hot, Word2Vec, and Global Vectors (GloVe) have the potential to capture this type of meaningful semantic relationship between pairs of words and/or syntactic information from unstructured text data. More recent language models encode different vectoral representations for a word depending on its context, which is important for words with multiple meanings (i.e., polysemy) (Devlin et al., 2019; Sanh et al., 2020). Some of these more advanced contextual language models, such as the Bidirectional Encoder Representations from Transformers (BERT) algorithm, train a neural network to associate each word with every other word in a sentence (i.e., bi-directional self-attention) (Devlin et al.,

* Corresponding author.
E-mail address: christopher.lawley@nrcan-rncan.gc.ca (C.J.M. Lawley).

https://doi.org/10.1016/j.acags.2022.100084
Received 19 October 2021; Received in revised form 10 April 2022; Accepted 25 April 2022
Available online 4 May 2022

2019; Vaswani et al., 2017). The ability to encode bi-directional sequence order and long-range dependencies of words is important to address polysemy and more complicated sentence semantics, allowing contextual language models like BERT to outperform more simple, static, and non-contextual language models in several NLP tasks (Wang et al., 2019).

However, despite often being trained on vast digital corpora comprising billions of words, language models that are trained on general text are often missing the vocabulary and concepts that are required to make meaningful predictions in some scientific sub-domains (Tshitoyan et al., 2019). More recent NLP research has thus become focused on retraining language models with domain-specific text (Gururangan et al., 2020). A large number of domain-specific language models have been developed with improved understanding of the semantic information in their field of expertise, therefore leading to better performances on the domain-specific tasks, including BioBERT (Lee et al., 2019), E-BERT (Zhang et al., 2021), PatentBERT (Lee and Hsiang, 2019), SciBERT (Beltagy et al., 2019), and TweetBERT (Qudar and Mago, 2020). In contrast, language models that are specific to geoscience are rare, with the exception of some recent NLP downstream applications in translation (Gomes et al., 2021), keyword generation (Qiu et al., 2019a), information retrieval (Qiu et al., 2018), document search (Holden et al., 2019), and other forms of text mining (Enkhsaikhan et al., 2021a, 2021b; Ma et al., 2020; Peters et al., 2018; Wang et al., 2018). The results from this NLP research provide new tools for extracting geoscience knowledge from unstructured text, but tend to focus on evaluating language model performance on specific downstream tasks. For example, several recent NLP studies have focused on named entity recognition (NER), an important downstream task for naming and locating geoscientific properties from unstructured text (Enkhsaikhan et al., 2021a; Qiu et al., 2019b). The NER task requires training language models on large volumes of labelled data that have only recently become available in geoscience (Enkhsaikhan et al., 2021b). Instead, the current study: (1) trains word embeddings using unsupervised methods and unlabelled geoscientific text data; and (2) measures the quality of word embeddings using a suite of intrinsic evaluation criteria rather than tuning model performance for any single downstream task. Once evaluated, these pre-trained geoscience language models can be applied to a range of downstream tasks, as recently demonstrated by Fuentes et al. (2020) for 3D geological modelling.

Continued progress on geoscience-specific word embeddings is required given the important role that qualitative descriptions of the rock record have had on the development and application of the science. For example, billions of dollars are spent every year by the mineral exploration and mining industries to describe the lithology, mineralogy, colour, texture, structure, cross-cutting relationships, and other geological attributes of drill core. In most cases, these written rock descriptions are stored as unstructured text fields within core logging software and databases. Word embeddings can leverage these qualitative geological observations for applications such as rock classification and predictive modelling (Fuentes et al., 2020; Joshi et al., 2021). Unfortunately, most of the existing work on geoscience language modelling and word embeddings are trained on private company reports or peer-reviewed publications that require paid subscriptions (Bayraktar et al., 2019; Consoli et al., 2020; Gomes et al., 2021; Padarian and Fuentes, 2019; Qiu et al., 2019a). The few available published examples are also based on languages other than English (Consoli et al., 2020; Gomes et al., 2021; Ma et al., 2021) and, with the exception of Padarian and Fuentes (2019), the trained models are rarely published alongside the method description. Moreover, the criteria for evaluating the performance of geoscience word embeddings are rarely published (Padarian and Fuentes, 2019).

This study addresses each of those issues and makes the following contributions: 1) we present two geoscience-specific language models using the GloVe and BERT methods, which are trained on public geoscientific documents written in English. The required text data from government geoscientific reports are extracted, processed for consistent formatting, and combined with a subset of open access and peer-reviewed publications; 2) we train a new geoscience-specific tokenizer, which is the method used by BERT for breaking words into sub-words, or tokens, to improve performance for geoscientific text; 3) we present four geoscience-specific evaluation tasks. These intrinsic evaluation criteria (i.e., analogy, clustering, relatedness, and nearest neighbours) address the quality of the word embeddings for capturing meaningful semantic relationships and are based on commonly used metrics in previously published NLP research (Mikolov et al., 2013a, 2013b; Padarian and Fuentes, 2019); 4) we further demonstrate the application of these geoscience language models using unsupervised machine learning to extract geochemical and mineral assemblages from the word embedding space for the first time; and 5) we release the NRCan training text dataset, language models, evaluation tasks, and the source code to the community.

2. Language modelling data and methods

2.1. Text datasets

The text datasets used in the current study contain a variety of geoscientific publications sourced from the Natural Resources Canada (NRCan) GEOSCAN publications database (n = 27,081 documents), provincial government publication databases (e.g., Ontario Geological Survey, Alberta Geological Survey, and British Columbia Geological Survey; n = 13,898 documents), and a subset of open access journals (e.g., Materials, Solid Earth, Geosciences, Geochemical Perspective Letters, and Quaternary) available through the Directory of Open Access Journals (DOAJ; n = 3998 documents) (Fig. 1; Table 1). Scanned government publications were pre-processed to remove figures, maps, tables, references, and other irregularly formatted text prior to analysis. Artifacts generated from optical character recognition (OCR) and low-quality scanned PDFs s from the GEOSCAN publications database were also excluded from further analyses (i.e., the total GEOSCAN database contains approximately 83 k documents; however a much smaller subset were readily available for use as part of the current study). Texts were extracted from high-quality pdf documents using "pdfminer" (https://github.com/euske/pdfminer) and converted to delimited text files for further analysis. The pre-processing steps applied were: 1) removing punctuation; 2) replacing upper casing; 3) converting all non-ascii characters to their ascii equivalent; 4) removing non-printable characters; 5) splitting and adding space around punctuation; 6) removing new lines; 7) removing title pages and/or tables of contents; 8) removing specific forms of alpha-numeric data (e.g., DOIs, URLs, emails, and phone numbers); 9) removing French text; 10) merging split words; 11) filtering text boxes that contain an insufficient percentage (80–90%) of detectable words; and 12) merging all of the extracted text for each document. For the BERT model discussed below, sentence tokenization was completed using the "en_core_web_lg" language model included with spaCy library (https://spacy.io/). The data pre-processing source code is made freely available as part of the current study (https://github.com/NRCan/geoscience_language_models). Text data from the GEOSCAN publications database are freely available from Raimondo et al. (2022).

2.2. Non-contextual language modelling (GloVe)

Words that occur together tend to be more closely associated (Mikolov et al., 2013a; 2013b). Examples of this co-occurring relationship in the context of geoscience, include minerals and their element constituents. Herein the GloVe method was used to map each word in the training corpus to a numerical vector in N-dimensional space (Pennington et al., 2014). Simple vector arithmetic can then be used to infer semantic relationships between pairs of words or quantitatively identify the nearest neighbours to words using either the Euclidean distance or

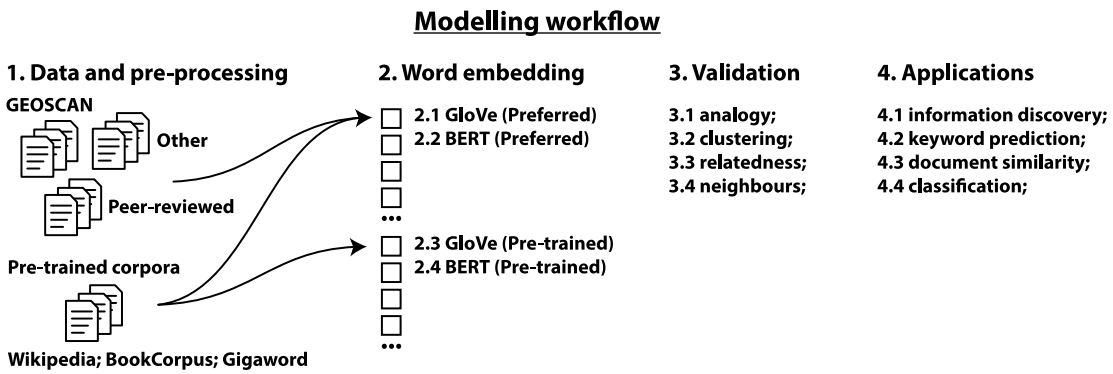

Fig. 1. Modelling workflow for the present study. Public geoscientific reports from multiple government publication databases and a subset of peer-reviewed publications were processed and combined as part of the current study to retrain previously published language models using geoscientific text (i.e., preferred GloVe and BERT). Pre-trained language models, in contrast, are based on a much larger, but general corpora (i.e., Wikipedia, Gigaword, BookCorpus). All four models are evaluated using a variety of geoscience-specific analogy, clustering, relatedness, and nearest neighbour tasks.

Table 1
Geoscientific datasets.

Data sources	Dataset	Publication counts (n)
Natural Resources Canada	GEOSCAN	27,081
Ontario Ministry of Energy, Northern Development and Mines	Publication database	10,614
Alberta Geological Survey	Report Database	1011
British Columbia Geological Survey	Natural Resource Online Services	2273
Directory of open access journals (DOAJ)	Materials, Solid earth, Geosciences, Geochemical Perspectives Letters, Quaternary	3998
	Total geoscientific publications (n)[a]	44,977

[a] Geoscientific publications correspond to the following word and vocabulary counts for each model: GloVe words = 211 M; GloVe vocabulary = 1.76 M; BERT words = 350 M; BERT vocabulary = 3.62 M.

cosine similarity in vector space. The generic, or "pre-trained", GloVe model is based on a relatively large corpora taken from Wikipedia (2014) and the 5th Edition of English Gigaword (Parker et al., 2011), comprising billions of words or sub-words (Pennington et al., 2014). This pre-trained GloVe model was used as a baseline to evaluate whether continued retraining using a much smaller but domain-specific corpora could improve model performance (i.e., the preferred GloVe model). Both iterations of the GloVe model (i.e., pre-trained and preferred) were trained using AdaGrad (Duchi et al., 2011) with the most abundant tokens (i.e., minimum frequency of 5), considering a context window of size 15 for 15 iterations, fixed weighting functions (x_max = 10 and alpha = 0.75), and multiple vector dimensions (i.e., 50 d and 300 d) as described by Pennington et al. (2014). Models that used a relatively small number of vector dimensions tended to yield lower scores for the intrinsic evaluation criteria described below and thus all of the described GloVe models use 300-dimensional vectors.

2.3. Contextual language modelling (BERT)

Contextual language models consider words and their neighbours for a more complete representation of their meaning, and thus, unlike non-contextual methods, yield multiple representations for each individual word. The latest contextual language models, such as XLNet (Yang et al., 2020) and BERT (Devlin et al., 2019), represent some of the most popular NLP architectures and yield state-of-the-art performance on tasks included within the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2019). These models are based on a relatively complex multi-layer bidirectional transformer encoder architecture and are pre-trained on a giant corpus, which avoids having to train new language models from scratch for general NLP tasks. For example, the original BERT model was pre-trained on the BookCorpus dataset (Zhu et al., 2015) and English Wikipedia, comprising billions of words. Herein we further train the DistilBERT model (Sanh et al., 2020), which achieves similar performance to the original BERT method but yields smaller models that are easier to train, less susceptible to overfitting, and are more appropriate for smaller datasets. The approach of updating the existing DistilBERT model by further training using the masked-language-modelling objective is distinct from training the preferred GloVe model from scratch.

First, pre-processed text is converted to tokens that may include words, sub-words or punctuation. Sub-word tokenization limits the number of out-of-vocabulary words, which allows BERT models trained on general corpora to be applied to specific sub-domains. However, tokenization must be applied consistently during model training and inference and the impact of this process on geoscience-specific words has not been previously evaluated. Two different tokenization methods were tested as part of the current study: (1) the original pre-trained WordPiece tokenizer for BERT (Devlin et al., 2019); and (2) multiple geoscience-specific tokenizers that were created by adding geoscience tokens prior to continued pre-training using the geoscientific corpora (i.e., the preferred BERT model). The geoscience tokens were identified by training the WordPiece tokenizer on the same geoscientific corpora. The original BERT tokenizer vocabulary has 994 unused "blank" tokens, such that the associated model weights remain at their original randomly initialized values. More tokens tend to yield more complete words, and, based on our evaluation, substituting 250 of those unused tokens with geoscience-specific tokens tended to produce the best-performing language models using the intrinsic evaluation criteria below. However, BERT models are typically evaluated based on their performance on downstream NLP tasks using complete sentences or whole paragraphs (Discussed below). In order to validate these BERT models using the intrinsic, geology-specific evaluation criteria, individual words were converted to a numeric representation using the final layer's vector with only the individual words used as input. For words represented by multiple tokens, the average of the final layer vectors for those sub-words was used for the purposes of intrinsic evaluation. The pre-trained and preferred BERT (i.e., using the geo-tokenizer and geoscientific corpora) models were generated using the "HuggingFace" machine learning library (https://huggingface.co/) (Wolf et al., 2020) with the same combination of hyperparameters (e.g., learning rate = $5e^{-5}$ and $2.5e^{-5}$; batch size = 48; max steps = 1 and 3 million; warm-up steps: 0, 100 k, 300 k) described in the original Devlin et al. (2019) method.

3. Intrinsic evaluation methods and results

3.1. Analogies

Analogy tests have been extensively used to evaluate whether the representations of language models capture structural and syntactic similarities between words (Mikolov et al., 2013a, 2013b; Pennington et al., 2014). However, to test whether the representations capture geoscience-specific concepts, we require geology-specific analogies, yet no such standard test-suite exists nor have previous geology studies released their analogy suites (Padarian and Fuentes, 2019). Thus, following Padarian and Fuentes (2019), we develop a suite of geoscience-specific analogy quartets to test whether geoscience concepts are captured by word vectors (n = 55) (Electronic Supplementary Material Table 1). For example, basalt (a) is to mafic (b) as rhyolite (x) is to felsic (y) can be evaluated as four separate linear equations in vector space:

$$x + b - y = a$$

$$a + y - x = b$$

$$a + y - b = x$$

$$x + b - a = y$$

Each combination of the analogy quartet was evaluated in turn using the rank statistic and cosine distance to the "correct" answer (i.e., a measure of directional similarity) in the vector embedding space. Analogy quartets that yield smaller cosine directions, or lower rank, to the "correct" answer represents a relative indicator of good model performance. The average of these metrics for each of the analogy quartets that are present within the model vocabularies are presented in Fig. 2 and reported in Table 2. Overall, the preferred GloVe and BERT language models outperform their pre-trained versions for virtually all geoscience-specific analogies (Fig. 2). Model improvements are most significant for the preferred GloVe model, which yields a 25% improvement (i.e., smaller cosine distance) on the analogy task over its pre-trained version. The preferred GloVe also typically predicts the correct answer within ten of the most closely associated words based on the rank statistic (Fig. 2b). Model improvements for the preferred BERT models, in contrast, are only slightly better than the pre-trained version (3% improvement; Fig. 2c). The analogy task further suggests that the

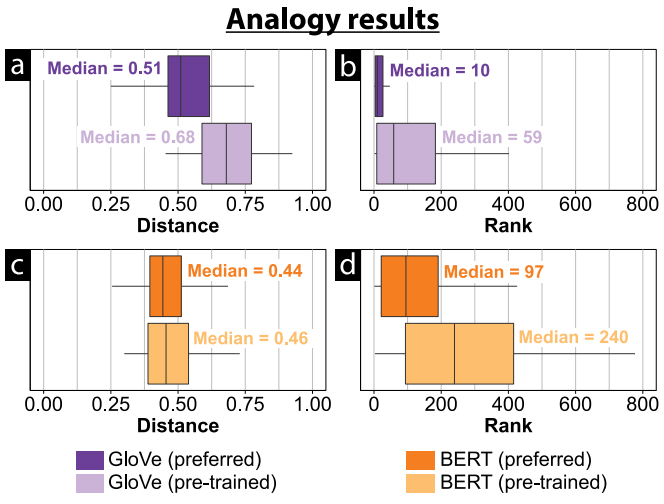

Fig. 2. Analogy results for the pre-trained and preferred GloVe and BERT language models. Results suggest that the language models trained on geoscientific text tend to outperform their pre-trained versions (i.e., lower rank and distance). The preferred GloVe and BERT language models yields the lowest analogy rank and distance for the analogy task, respectively.

Table 2
Intrinsic evaluation results.

Task	Geoscience language models			
	pre-trained GloVe	preferred GloVe	pre-trained BERT	preferred BERT
Analogy (median rank)	61	23	240	97
Analogy (median distance)	0.684	0.538	0.455	0.443
Clustering (median score)	0.814	0.900	0.758	0.797
Relatedness (Pass %)	86%	92%	69%	78%

preferred GloVe model outperforms the preferred BERT model by approximately 13% (Fig. 2a, c).

3.2. Clustering

Closely associated words tend to group together in vector space and the clustering of groups of words provides a second intrinsic evaluation metric for assessing language model performance. Clusters of words are based on the GeoSciML and/or EarthResourceML vocabularies (www.geosciml.org) (Raymond et al., 2012; Sen and Duffy, 2005; Simons et al., 2006). These standard vocabularies were previously grouped into 16 categories, or clusters, by subject matter experts contributing to the International Union of Geological Sciences (IUGS) Commission for the Management and Application of Geoscience Information (CGI; i.e., Alteration type, Commodities, Compositions, Environment, Environmental impact, Events, Exploration activity, Fault types, Foliation types, Genetic, Geometry, Lineations, Metamorphic facies, Particle shapes, Particle types, and Rock types). Simple Naïve Bayes models (Chan et al., 1982) were then trained using leave-one-out-classification for each word in each of the previously defined GeoSciML and EarthResourceML clusters. This process was repeated for every possible pair of clusters to test whether the word embedding space for each language model could be divided into groups of words with semantic similarities. The median and distribution of classification scores for each cluster and model are presented in Fig. 3 and reported in Table 2. Overall, the clustering task demonstrates that the preferred GloVe (median = 0.90) and BERT (median = 0.80) models tend to outperform their pre-trained versions (GloVe = 0.81 and BERT = 0.76) for each of the previously published clusters. Classification results further suggest that the preferred GloVe model yield the best performance for each of the 16 categories included within this clustering task.

3.3. Relatedness

Relatedness is a separate form of intrinsic evaluation that is based on pairs of similar words. Ideally, words with similar meaning yield closely related vector representations that are, in turn, dissimilar to unrelated words. Our test suite for the relatedness evolution comprises 249 pairs of similar words in 12 themes proposed from subject matter experts as part of the current study (Electronic Supplementary Material Table 1). Each similarity pair was then exhaustively assigned "dissimilar" words from the same (i.e., intra-theme) and different (i.e., outra-theme) themes. World triples were then evaluated as a "Pass" or "Fail" depending on whether the similarity pair yielded a lower cosine distance than the dissimilar word (Fig. 4). Language models that correctly predict more of these similarity pairs are considered to be relatively effective at identifying words with similar meaning. Overall, the preferred GloVe (92% Pass) and BERT (78% Pass) language models outperform their pre-trained versions on the relatedness task (GloVe = 86% Pass; BERT = 69% Pass; Table 2). Model performance for a subset of the most common themes is presented in Fig. 4. Overall, the preferred GloVe model yields the best pass rate for the relatedness task (Fig. 4).

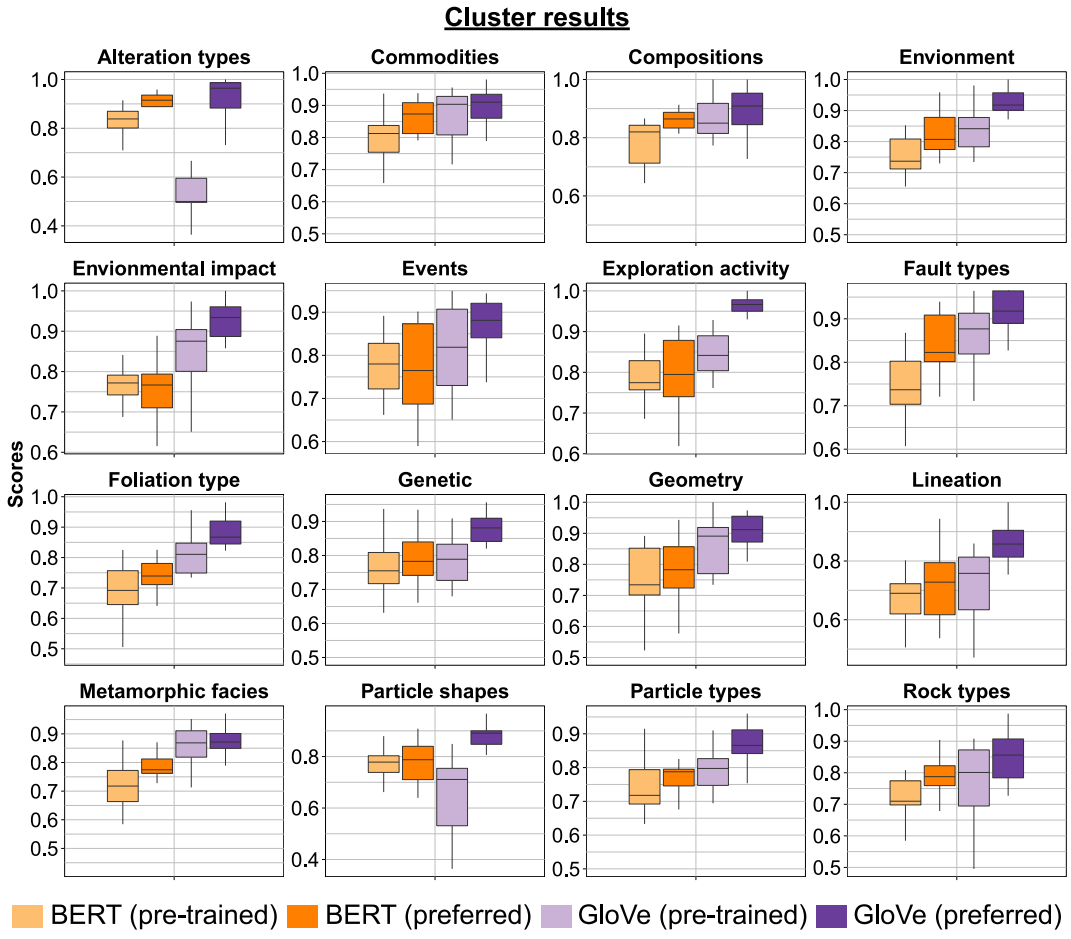

Fig. 3. Cluster results for the pre-trained and preferred GloVe and BERT models. Each cluster category was previously defined and are based on the GeoSciML and EarthResourceML vocabularies (Sen and Duffy, 2005). Results suggest that language models trained on geoscientific text tend to outperform their pre-trained versions. The preferred GloVe language model yields the best performance for the clustering task.

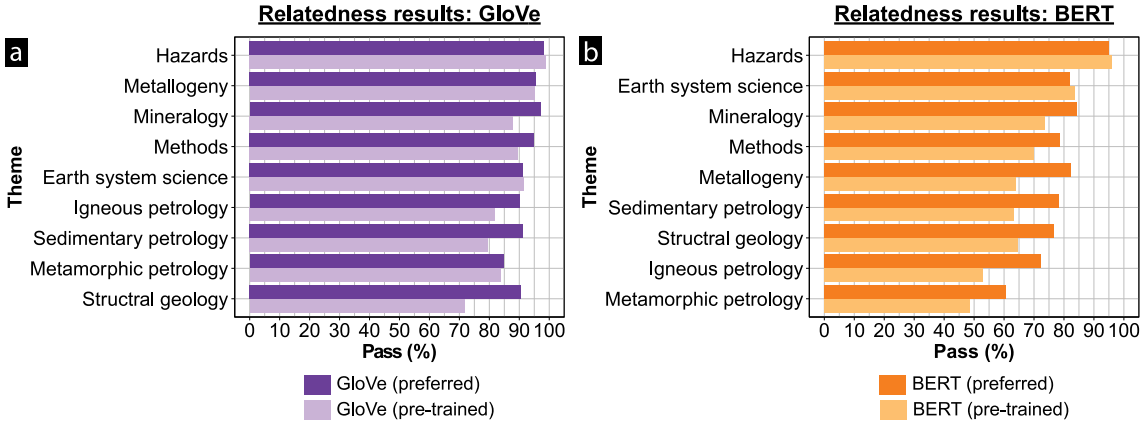

Fig. 4. Relatedness results for the pre-trained and preferred GloVe (a) and BERT models (b). Relatedness is based on whether word embedding for similarity pairs are closer than dissimilar words for a subset of the most common themes. Results suggest that language models trained on geoscientific text tend to outperform their pre-trained version, particularly for themes with more domain-specific vocabularies (hazards versus petrology). Overall, the preferred GloVe model yields the best performance for the relatedness task.

3.4. Nearest neighbours

A qualitative assessment on the appropriateness of nearest neighbours represents the fourth form of intrinsic evaluation included as part of the current study. Five words representing the range of research being conducted at the Geological Survey of Canada were selected for this task (i.e., Earth, Exploration, Environment, Climate, and Hazard). All four language models yield a reasonable set of the ten nearest neighbours to each word. However, the nearest neighbours for the pre-trained GloVe and BERT models tended to be more general in nature and less focused on geoscience research specifically. The "Hazard" and "Earth" categories provide the starkest example of this effect (Table 3 and Table 4), with

Table 3
Nearest neighbour results for preferred and pre-trained GloVe models.

Earth	Earth (pre-trained)	Exploration	Exploration (pre-trained)	Environment	Environment (pre-trained)	Climate	Climate (pre-trained)	Hazard	Hazard (pretrained)
sciences	planet	prospecting	explorations	environments	environments	climatic	warming	hazards	hazards
physics	mars	drilling	drilling	environmental	environmental	warming	environment	risk	danger
science	planets	mining	prospecting	conditions	climate	change	global	mitigation	risk
journal	orbit	programs	explore	depositional	sustainable	impacts	change	vulnerability	pose
planetary	moon	diamond	exploring	deposition	development	adaptation	environmental	risks	dangers
sci	spacecraft	discovery	offshore	marine	ecology	warmer	climatic	earthquake	posed
v	martian	development	mining	impacts	ecological	global	climates	probabilistic	poses
crust	universe	companies	discovery	setting	conditions	ecosystems	biodiversity	threat	risks
planet	space	explored	exploratory	settings	protection	changing	weather	landslides	safety
evolution	orbiting	discoveries	discoveries	nature	biodiversity	changes	greenhouse	landslide	contamination

Table 4
Nearest neighbour results for preferred and pre-trained BERT models.

Earth	Earth (pre-trained)	Exploration	Exploration (pre-trained)	Environment	Environment (pre-trained)	Climate	Climate (pre-trained)	Hazard	Hazard (pre-trained)
planet	planet	drilling	drilling	climate	ecology	weather	weather	impact	disturbance
moon	space	evaluation	excavation	ecology	climate	environment	environment	landslide	alteration
human	moon	reconnaissance	reconnaissance	life	sustainable	precipitation	precipitation	earthquake	seismic
life	life	diamond	extraction	health	resource	cooling	geology	tsunami	intrusion
crust	soil	research	observation	ocean	earth	temperate	soil	flood	asbestos
contemporary	gravity	petroleum	evolution	habitat	life	runoff	cooling	trajectory	prediction
environment	environment	sampling	research	evolution	soil	feedback	weathering	instability	radioactive
discus	human	zoning	erosion	contemporary	energy	arid	wind	hurricane	sedimentary
wasting	surface	extension	spectroscopy	zoning	health	heat	radiation	runoff	enveloped
enveloped	terrane	alteration	assessment	estuary	water	cold	tropical	shattering	strata

pre-trained models predicting words that are more loosely associated with space (e.g., “mars”, “martian”, “spacecraft”, “moon”) for the “Earth” category. Differences between the preferred and pre-trained models for the other categories are more difficult to evaluate given the large amount of semantic overlap for the closest matching words. Nearest neighbour results for the two preferred language models are also similar, except perhaps for the “Earth” and “Environment” categories, which were slightly more appropriate for the preferred BERT model.

3.5. Principal component analysis

Principal component analysis is an unsupervised technique that can be used to reduce high dimensional word embedding vectors for further visualization and analysis (Mikolov et al., 2013a; 2013b). Herein principal component analysis was calculated using the 300 dimensional vectors from the preferred GloVe model and the “prcomp” function in R (R Core Team, 2021). Words that occur together and/or share some semantic relationship tend to plot close together in principal component space (Mikolov et al., 2013a; 2013b) and can be used to organize subsets of words along gradients without supervision (Fig. 5). Words at the end of each gradient are the most dissimilar and their correct order along the first principal component tends to suggest that their semantic differences represent the largest source of dataset variance. For example, the first principal component correctly orders sedimentary grain sizes (i.e., coarse gravel to fine mud; Fig. 5a), igneous intrusive rock compositions (i.e., ultramafic harzburgite, mafic gabbro, intermediate granodiorite, and felsic granite; Fig. 5b), and metamorphic grade (i.e., granulite to zeolite facies; Fig. 5c). These principal component analysis results are similar to some of the analogies described in Padarian and Fuentes (2019).

Alternatively, multiple principal components can be visualized together on a principal component analysis biplot to identify the multivariate relationships between word quartets and their analogies. Multiple analogy examples are presented in Fig. 6 to demonstrate that pairs-of-words plot in distinct quadrants for each biplot. For example, rocks and their corresponding grain size (Fig. 6a), minerals and their

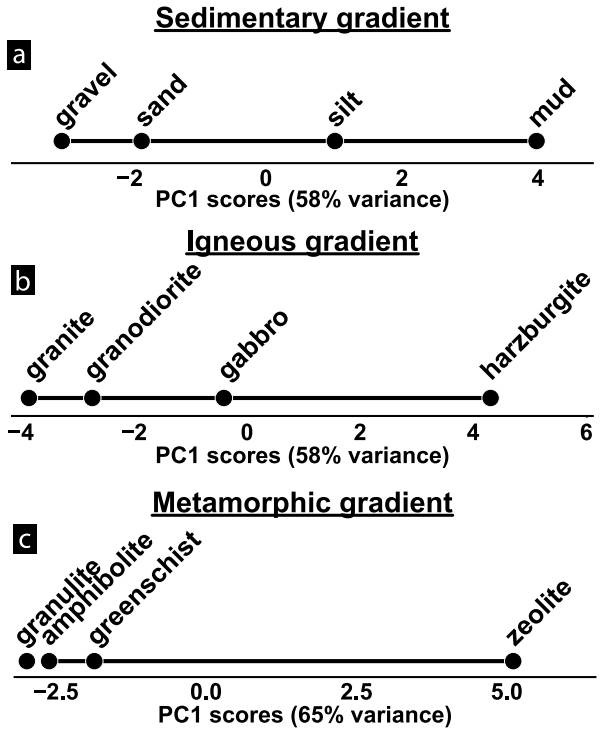

Fig. 5. (a–c) Word vectors that are relevant to sedimentary (a), igneous (b), and metamorphic petrology re-calculated after principal component analysis (PCA) and plotted along the first principal component (PC1). The correct order of words in PCA space suggests that the GloVe-based word embedding encode meaningful semantic relationships, which, in this case, reflect well-known rock classification schemes.

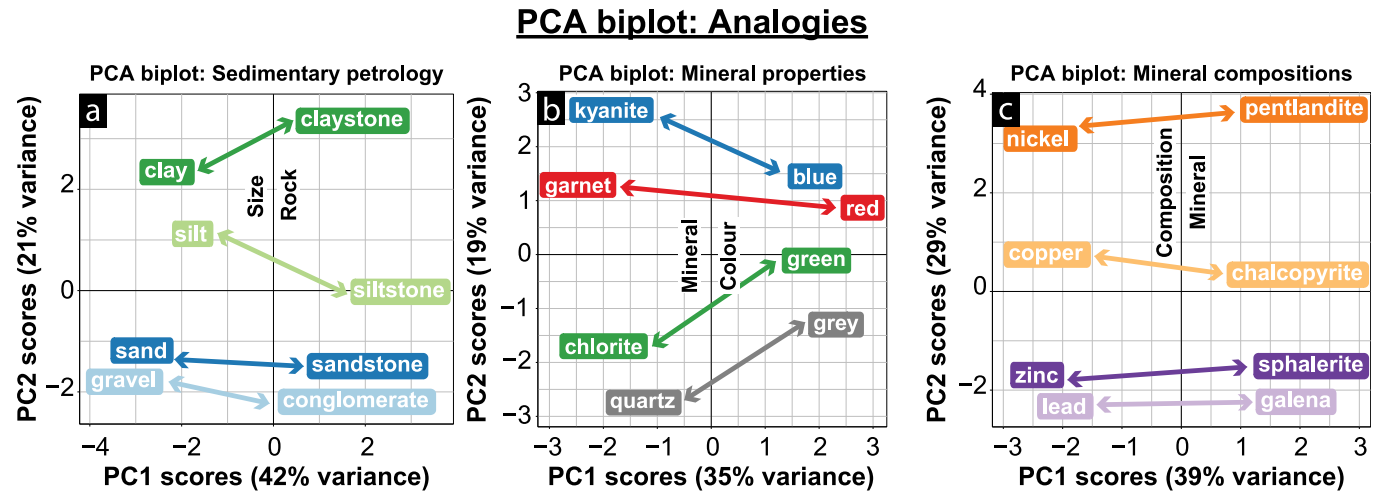

Fig. 6. (a–c) Principal component analysis (PCA) biplots showing the relationships between multiple word analogies. Words with similar semantic relationships cluster together in PCA space and define gradients along different principal component axes (PC1 and PC2). Connections between sediment grain size and rock type, (a) minerals and their properties (b), and minerals and their chemical constituents (c) suggest that word embeddings are capable of capturing meaningful semantic relationships that can be used in other natural language processing tasks.

properties (Fig. 6b), and minerals and their constituent elements (Fig. 6c) are divided into positive and negative PC1 scores for each biplot, suggesting that the underlying concept for each analogy represents the largest source of variance. The correct ordering of grain size for detritus and rock types along the second principal component (i.e., PC2) further suggests that multiple concepts can be visualized simultaneously, providing evidence for more complex concepts preserved from linear transformation of the preferred GloVe model vectors.

More complex concepts are presented in Figs. 7 and 8, which show the linearly transformed word vectors colour coded to their geochemical behaviour (Goldschmidt, 1937) and mineralogy (Gaines et al., 1997), respectively. All elements in the periodic table were searched and matched where possible (n = 96) using their full names in word embedding space prior to plotting with their abbreviated form for the purposes of visualization (Fig. 7). The clustering of element names with similar Goldschmidt classifications (e.g., Rare Earth Elements) suggests that word embeddings preserve some of the physical and chemical properties of individual elements (e.g., atomic radius, bonding characteristics, and/or electron configuration).

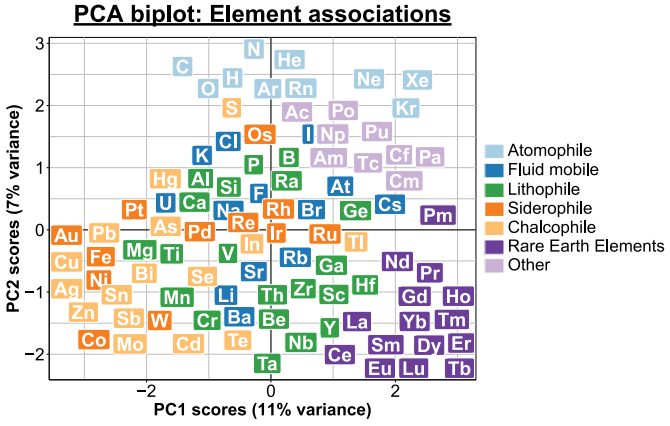

Fig. 7. Principal component analysis (PCA) biplot showing element abbreviations and colour coded to the Goldschmidt rules of geochemical behaviour (Goldschmidt, 1937). Full element names were searched in the word embedding space and were abbreviated for the purposes of visualization. Elements with similar geochemical behaviour tend to cluster together in PCA space and likely reflect the same mineralogical control on geochemistry observed in nature. However, in this case, the geochemical behaviour of elements is emergent from thousands of individual observations and their vectoral representation in word embedding space.

The well-known Dana classification scheme (Gaines et al., 1997) was used to test whether unsupervised learning can be used to provide additional intrinsic validation of the preferred GloVE model (Fig. 8). The original Dana classification subdivides over 4000 mineral species into at least 10 classes according to their composition and structure, with smaller subdivisions based on crystal symmetry (Gaines et al., 1997). Mineral names were searched and matched (number of matches = 1893) to their respective compositional groups defined by the Dana classification scheme (Gaines et al., 1997). The principal component analysis biplot reveal clear differences between the vectoral representations of disparate mineral classes that are broadly consistent with the mineral assemblages that occur in nature (Fig. 8).

4. Discussion

4.1. Intrinsic evaluation of geoscience language models

The new intrinsic evaluation criteria presented herein demonstrate that continued retraining of language models improves model performance on geology-specific tasks (Figs. 2–4; Table 2; Electronic Supplementary Material Table 1). These kinds of intrinsic evaluation criteria (i.e., analogies, relatedness, clustering, and nearest neighbours) attempt to evaluate the content and quality of the embeddings themselves, as opposed to their ability to improve performance on a downstream task. We interpret the relatively poor performance of the pre-trained language models on the geoscience-specific tasks as likely due to the limited frequency of domain-specific words in the general corpora despite the vast amount of text available in Wikipedia, Gigaword, and BookCorpus datasets. Non-contextual language models are particularly sensitive to these infrequent or out-of-vocabulary words because they are based on a static tokenizer (i.e., words are broken by white space, new lines, and punctuation), which coupled with the specific usage of words in geology, provide the most likely explanations for the 25% improvement observed for the preferred GloVe models over its pre-trained version for the analogy task (Fig. 2).

Improvements were also observed for the preferred BERT models over its pre-trained version (Figs. 2–4). Contextual language models like BERT are less sensitive to infrequent or out-of-vocabulary words because they are based on a combination of words and sub-words (i.e., tokens). Instead, we suggest that the improved performance of the preferred

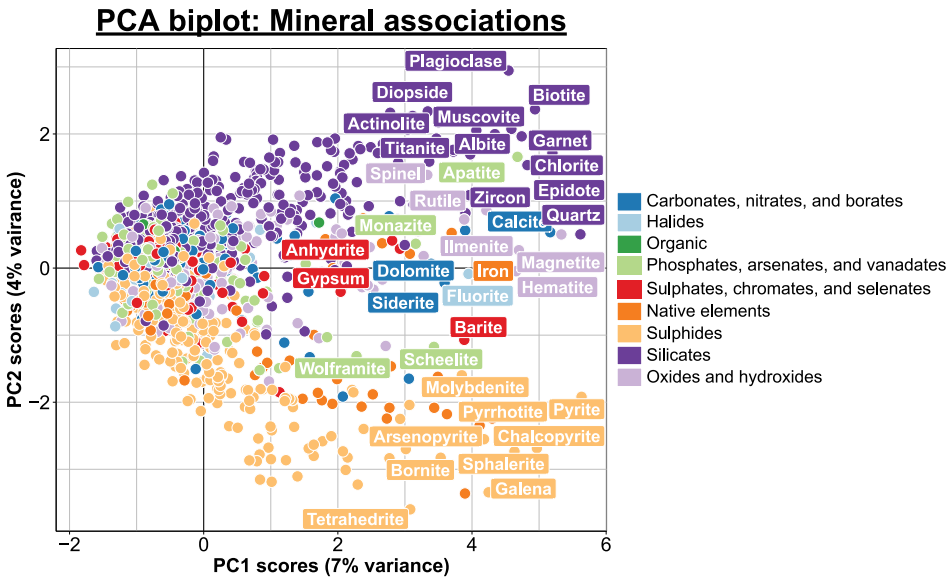

Fig. 8. Principal component analysis (PCA) biplot of minerals names and colour coded to the well-known Dana classification scheme (Gaines et al., 1997). Minerals with similar classifications plot together in PCA space, reflecting similar vector properties (e.g., silicates versus sulphides). Word embeddings provide a powerful framework for evaluating and predicting mineral groups based on thousands of observations in nature from multiple trained observers over time. Minerals from disparate classification groups that plot close together provide intriguing evidence for associations that require re-examination (e.g., the lesser known association between scheelite and molybdenite in porphyry-skarn mineral systems).

BERT model may be related to more meaningful sub-words generated from the geo-tokenizer. This hypothesis is supported by testing the performance of BERT models using different geo-tokenizers relative to the standard WordPiece tokenizer. For example, the pre-trained WordPiece tokenizer converts the word "seismology" to se + ism + ology; whereas the preferred geo-tokenizer identified through testing converts the same word to seism + ology. As discussed above, the preferred geo-tokenizer used in the preferred BERT model added 250 domain-specific tokens prior to continued retraining of the model. Some of these new geo-tokens that are likely significant for the present study, include "geolog", "formation", "atigraph", and "orph". Our approach of adding specific geo-tokens is appropriate for the relatively small number of public geoscientific text harvested from government and open-access peer-reviewed publications. However, it is possible that continued retraining of BERT models with a larger geoscientific corpora could further improve the performance on the intrinsic evaluation criteria. In general, smaller models with a larger domain-specific training dataset reduces the chance of overfitting and improves performance.

The improved performance of the new geological language models relative to their pre-trained version add to a growing number of studies that point to the benefit of continued retraining for domain-specific tasks. For example, Padarian and Fuentes (2019) report that their GeoVec word embedding improved the accuracy for their intrinsic evaluation tests by 108% relative to the pre-trained GloVe model. The GeoVec model was trained on a much larger suite of full-text geoscientific articles harvested from the Elsevier ScienceDirect application programming interface (n = 280,764). Similarly, Bayraktar et al. (2019) and Ma et al. (2021) suggest that continued retraining on geoscience data can improve performance for their BERT models in downstream NLP tasks (e.g., document summary and similarity). These studies represent some of the few published examples comparing generic and domain-specific language models in geoscience. Future research should also consider evaluating geoscience language models using other forms of intrinsic (e.g., perplexity) and extrinsic evaluation criteria (e.g., NER). For example, Consoli et al. (2020) and Qiu et al. (2019a) tested word embeddings retrained geoscientific documents using NER in Portuguese and Chinese, respectively. Extrinsic evaluation following this approach require a large number of documents that have been manually annotated by experts and the scarcity of labelled geoscientific documents presents a number of challenges for testing language model performance using NER (Enkhsaikhan et al., 2021b). Ideally, geoscience language models should be evaluated using extrinsic criteria that are specific to their particular use. This extra stage of validation during application is important given that language model performance is expected to vary for different forms of evaluation, as documented by NLP research in other domains (Santos et al., 2020; Wang et al., 2019).

4.2. *Knowledge extraction and other language model applications*

Advances in machine learning and NLP are providing new numerical tools for downstream predictive text applications and/or have the potential to unlock the hidden information within unstructured data (Bengio et al., 2000; Hirschberg and Manning, 2015). New research results presented above provide additional support for the future of these NLP methods in geoscience. For example, the gradients between the composition (i.e., mafic to felsic), grain size (i.e., coarse to fine) and metamorphic grade (i.e., high to low) of rock types observed after translating word vectors to principal components, demonstrates that the preferred GloVe model can be used to predict the correct order of words according to well-established geological concepts without any supervision for keyword generation, summarization, and translation tasks (Figs. 5 and 6). Fuentes et al. (2020) further demonstrate that simple, non-contextual word embeddings can also be used as input into predictive models that classify rocks based on their geological descriptions. Given that billions of dollars are spent on these types of drilling campaigns every year, free NLP methods that are built on public geological word embeddings have great potential to maximize the return on this investment.

The potentially more challenging application of word embeddings is the discovery of latent information that may be stored within the vast volumes of unstructured geological text. Examples of knowledge discovery from word embeddings are still relatively rare, although Tshitoyan et al. (2019) demonstrated that new materials could have been proposed years before their first reporting from information buried within the scientific literature. The discovery of this type of new information from the published literature is possible because language models that are based on unsupervised learning of massive text datasets (i.e., annotated training data are not required) are likely to find connections and patterns of research results that may have been difficult to identify manually (Ma et al., 2017). The clustering of elements with similar geochemical behaviours, as originally described by Goldschmidt (1937), in PCA space highlights the potential for similar knowledge discovery to occur in unstructured geoscientific text (Fig. 7). We interpret the clusters of elements with similar geochemical behaviour in

principal component analysis space to reflect the underlying geological controls of natural processes since very few geoscientific publications would report results for all of these elements at the same time. This result is somewhat surprising given that word embeddings represent the accumulated signal from a large number of authors who have described these chemical associations independently and in their own words for over a century (i.e., GEOSCAN publications date back to 1845). Major and minor element substitution reactions are particularly apparent after the linear transformation of word vectors, as suggested by the clustering of major and minor element pairs (e.g., iron and nickel). Lesser known geochemical behaviour, such as the Rare Earth Elements characteristics of some lithophile elements (e.g., yttrium), highlight that multivariate statistical analysis of the linearly-transformed vectors can be further applied to re-classify elements that exhibit transitional behaviours in different geological environments.

Mineral associations represent one of the other important methods for tracing these geological environments back in deep time. New unsupervised machine learning results demonstrate that mineral groups yield word vectors that cluster together in multivariate space, which can be used to predict new mineral associations that may or may not have already been observed in nature. For example, magnetite and hematite are both iron-bearing minerals that represent important tracers of oxidation and reduction reactions and closely cluster with native iron (Fig. 8). Mineral associations such as these can be thus used to infer paleo-environmental conditions, track geological process through time, explore for new mineral resources, and predict the most favourable settings for so-called "missing" minerals (Hystad et al., 2019; Morrison et al., 2017, 2020). However, unlike manually curated mineralogical databases (Hazen, 2014; Morrison et al., 2017), we demonstrate herein that word embeddings capture at least some of these mineral associations from unsupervised machine learning of unstructured text. Combining these mineral names with other known mineral properties and using other multivariate statistical methods to characterize more subtle mineral associations represent important areas of future research. Our results add to the growing body of literature focusing on such data-driven discovery within geoscience (Hazen, 2014; Ma et al., 2017; Peters et al., 2018).

5. Conclusion

Recent advances in machine learning and NLP, coupled with the increased availability of high-performance computing in the cloud, are providing new tools to extract knowledge from the vast volumes of unstructured text. However, generic language models trained on general corpora are likely missing some of the specialized words and concepts that are specific to the sciences, suggesting that continued re-training with domain-specific text has the potential to improve model performance. Herein we apply some of these latest NLP tools to develop geoscience-specific language models. We demonstrate that contextual and non-contextual language models trained on geoscientific publications outperform the generic and pre-trained models on NLP tasks that are specific to geosciences. Whilst the relatively simple and non-contextual GloVE models yielded the best results on these specific tasks, more advanced contextual language models such as BERT are likely better for performing downstream NLP applications (e.g., sentiment analysis, keyword prediction, classification) and capture the more complete meaning of phrases and sentences. Nevertheless, we demonstrate how non-contextual word embeddings can be used on their own to make predictions from otherwise unstructured geological descriptions (Fuentes et al., 2020), and how the embedding space can be explored by statistical methods to highlight a number of features that are likely of significant interest (e.g., element associations and mineral assemblages).

Authorship statement

All authors: Conceptualization, Methodology and Writing; SR, TC, and AZ: Software, Resources, Data Curation; SR, TC, LB, AZ, and CJML: Formal Analysis, Investigation, and Validation.

Data availability

Raw and pre-processed text from the GEOSCAN publications database and the preferred GloVe and BERT language models are freely available from Raimondo et al. (2022).

Computer code availability

Language models were generated in python using open source libraries, including the HuggingFace Transformers library (Wolf et al., 2020). The source code used to load and manipulate word vectors is freely available from https://github.com/NRCan/geoscience_language_models.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was completed as part of the Targeted Geoscience Initiative program. Thanks to Boyan Brodaric and two anonymous reviewers for providing comments that improved this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.acags.2022.100084.

References

Bayraktar, Z., Driss, H., Lefranc, M., 2019. Representation learning in geology and GilBERT. In: Workshop on Document Intelligence at NeurIPS 2019, pp. 1–4.
Beltagy, I., Lo, K., Cohan, A., 2019. SciBERT: A Pretrained Language Model for Scientific Text arXiv preprint arXiv:1903.10676.
Bengio, Y., Ducharme, R., Vincent, P., 2000. A neural probabilistic language model. Adv. Neural Inf. Process. Syst. 13, 1–7.
Chan, T.F., Golub, G.H., LeVeque, R.J., 1982. Updating formulae and a pairwise algorithm for computing sample variances. In: Caussinus, H., Ettinger, P., Tomassone, R. (Eds.), COMPSTAT 1982 5th Symposium Held at Toulouse 1982. Physica-Verlag HD, Heidelberg, pp. 30–41. https://doi.org/10.1007/978-3-642-51461-6_3.
Chowdhary, K.R., 2020. natural language processing. In: Chowdhary, K.R. (Ed.), Fundamentals of Artificial Intelligence. Springer India, New Delhi, pp. 603–649. https://doi.org/10.1007/978-81-322-3972-7_19.
Consoli, B., Santos, J., Gomes, D., Cordeiro, F., Vieira, R., Moreira, V., 2020. Embeddings for Named Entity Recognition in Geoscience Portuguese Literature. LREC, pp. 4625–4630.
Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding arXiv:1810.04805 [cs].
Duchi, J., Hazan, E., Singer, Y., 2011. Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. 12, 2121–2159.
Enkhsaikhan, M., Holden, E.-J., Duuring, P., Liu, W., 2021a. Understanding ore-forming conditions using machine reading of text. Ore Geol. Rev. 135, 104200 https://doi.org/10.1016/j.oregeorev.2021.104200.
Enkhsaikhan, M., Liu, W., Holden, E.-J., Duuring, P., 2021b. Auto-labelling entities in low-resource text: a geological case study. Knowl. Inf. Syst. 63, 695–715. https://doi.org/10.1007/s10115-020-01532-6.
Fuentes, I., Padarian, J., Iwanaga, T., Willem Vervoort, R., 2020. 3D lithological mapping of borehole descriptions using word embeddings. Comput. Geosci. 141, 104516 https://doi.org/10.1016/j.cageo.2020.104516.
Gaines, R.V., Skinner, H.C.W., Foord, E.E., Mason, B., Rosenzweig, A., 1997. Dana's New Mineralogy: the System of Mineralogy of James Dwight Dana and Edward Salisbury Dana, eighth ed. Wiley-Interscience, New York.
Goldschmidt, V.M., 1937. The principles of distribution of chemical elements in minerals and rocks. The seventh Hugo Müller Lecture, delivered before the Chemical Society on March 17th, 1937. J. Chem. Soc. 655–673. https://doi.org/10.1039/JR9370000655.
Gomes, D. da S.M., Cordeiro, F.C., Consoli, B.S., Santos, N.L., Moreira, V.P., Vieira, R., Moraes, S., Evsukoff, A.G., 2021. Portuguese word embeddings for the oil and gas

industry: development and evaluation. Comput. Ind. 124, 103347 https://doi.org/10.1016/j.compind.2020.103347.
Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A., 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks arXiv:2004.10964 [cs].
Hazen, R.M., 2014. Data-driven abductive discovery in mineralogy. Am. Mineral. 99, 2165–2170. https://doi.org/10.2138/am-2014-4895.
Hirschberg, J., Manning, C.D., 2015. Advances in natural language processing. Science. https://doi.org/10.1126/science.aaa8685.
Holden, E.-J., Liu, W., Horrocks, T., Wang, R., Wedge, D., Duuring, P., Beardsmore, T., 2019. GeoDocA – fast analysis of geological content in mineral exploration reports: a text mining approach. Ore Geol. Rev. 111, 102919 https://doi.org/10.1016/j.oregeorev.2019.05.005.
Hystad, G., Morrison, S.M., Hazen, R.M., 2019. Statistical analysis of mineral evolution and mineral ecology: the current state and a vision for the future. Appl. Comput. Geosci. 1, 100005 https://doi.org/10.1016/j.acags.2019.100005.
Joshi, R., Madaiah, K., Jessell, M., Lindsay, M., Pirot, G., 2021. dh2loop 1.0: an open-source Python library for automated processing and classification of geological logs. Geosci. Model Dev. (GMD) 14, 6711–6740. https://doi.org/10.5194/gmd-14-6711-2021.
Lee, J., Yoon, W., Kim, Sungdong, Kim, D., Kim, Sunkyu, So, C.H., Kang, J., 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36, 1234–1240. https://doi.org/10.1093/bioinformatics/btz682.
Lee, J.-S., Hsiang, J., 2019. PatentBERT: Patent Classification with Fine-Tuning a Pre-trained BERT Model arXiv:1906.02124 [cs, stat].
Ma, K., Tian, M., Tan, Y., Xie, X., Qiu, Q., 2021. What is this article about? Generative summarization with the BERT model in the geosciences domain. Earth Sci. India. https://doi.org/10.1007/s12145-021-00695-2.
Ma, X., Hummer, D., Golden, J.J., Fox, P.A., Hazen, R.M., Morrison, S.M., Downs, R.T., Madhikarmi, B.L., Wang, C., Meyer, M.B., 2017. Using visual exploratory data analysis to facilitate collaboration and hypothesis generation in cross-disciplinary research. ISPRS Int. J. Geo-Inf. 6, 368. https://doi.org/10.3390/ijgi6110368.
Ma, X., Ma, C., Wang, C., 2020. A new structure for representing and tracking version information in a deep time knowledge graph. Comput. Geosci. 145, 104620 https://doi.org/10.1016/j.cageo.2020.104620.
Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013a. Efficient Estimation of Word Representations in Vector Space arXiv:1301.3781 [cs].
Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013b. Distributed Representations of Words and Phrases and Their Compositionality arXiv:1310.4546 [cs, stat].
Morrison, S.M., Buongiorno, J., Downs, R.T., Eleish, A., Fox, P., Giovannelli, D., Golden, J.J., Hummer, D.R., Hystad, G., Kellogg, L.H., 2020. Exploring carbon mineral systems: recent advances in C mineral evolution, mineral ecology, and network analysis. Front. Earth Sci. 208.
Morrison, S.M., Liu, C., Eleish, A., Prabhu, A., Li, C., Ralph, J., Downs, R.T., Golden, J.J., Fox, P., Hummer, D.R., Meyer, M.B., Hazen, R.M., 2017. Network analysis of mineralogical systems. Am. Mineral. 102, 1588–1596. https://doi.org/10.2138/am-2017-6104CCBYNCND.
Padarian, J., Fuentes, I., 2019. Word embeddings for application in geosciences: development, evaluation, and examples of soil-related concepts. SOIL 5, 177–187. https://doi.org/10.5194/soil-5-177-2019.
Parker, R., Graff, David, Kong, Junbo, Chen, Ke, Maeda, Kazuaki, 2011. English Gigaword Fifth Edition. https://doi.org/10.35111/WK4F-QT80.
Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543.
Peters, S.E., Husson, J.M., Czaplewski, J., 2018. Macrostrat: a platform for geological data integration and deep-time Earth crust research. G-cubed 19, 1393–1409. https://doi.org/10.1029/2018GC007467.
Qiu, Q., Xie, Z., Wu, L., Li, W., 2019a. Geoscience keyphrase extraction algorithm using enhanced word embedding. Expert Syst. Appl. 125, 157–169. https://doi.org/10.1016/j.eswa.2019.02.001.
Qiu, Q., Xie, Z., Wu, L., Li, W., 2018. DGeoSegmenter: a dictionary-based Chinese word segmenter for the geoscience domain. Comput. Geosci. 121, 1–11. https://doi.org/10.1016/j.cageo.2018.08.006.
Qiu, Q., Xie, Z., Wu, L., Tao, L., Li, W., 2019b. BiLSTM-CRF for geological named entity recognition from the geoscience literature. Earth Sci. India 12, 565–579. https://doi.org/10.1007/s12145-019-00390-3.
Qudar, M.M.A., Mago, V., 2020. TweetBERT: A Pretrained Language Representation Model for Twitter Text Analysis arXiv:2010.11091 [cs].
R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.
Raimondo, S., Chen, T., Zakharov, A., Brin, L., Kur, D., Hui, J., Burgoyne, S.L., Newton, G., Lawley, C.J.M., 2022. Datasets to support geoscience language models. Geological Survey of Canada. Open File 8848, 1 .zip file.
Raymond, O., Duclaux, G., Boisvert, E., Cipolloni, C., Cox, S., Laxton, J., Letourneau, F., Richard, S., Ritchie, A., Sen, M., Serrano, J.-J., Simons, B., Vuollo, J., 2012. GeoSciML v3.0 - a Significant Upgrade of the CGI-IUGS Geoscience Data Model, p. 2711.
Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2020. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter arXiv:1910.01108 [cs].
Santos, J., Consoli, B., Vieira, R., 2020. Word embedding evaluation in downstream tasks and semantic analogies. In: Proceedings of the 12th Language Resources and Evaluation Conference. Presented at the LREC. European Language Resources Association (ELRA), Marseille, pp. 4828–4834.
Sen, M., Duffy, T., 2005. GeoSciML: development of a generic GeoScience markup language. Comput. Geosci. Appl. XML Geosci. 31, 1095–1103. https://doi.org/10.1016/j.cageo.2004.12.003.
Simons, B., Boisvert, E., Brodaric, B., Cox, S., Duffy, T.R., Johnson, B.R., Laxton, J.L., Richard, S., 2006. GeoSciML: enabling the exchange of geological map data. ASEG Extended Abstracts 2006, 1–4. https://doi.org/10.1071/aseg2006ab162.
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, 95–98. https://doi.org/10.1038/s41586-019-1335-8.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention Is All You Need arXiv:1706.03762 [cs].
Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R., 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding arXiv:1804.07461 [cs].
Wang, C., Ma, X., Chen, Jianguo, Chen, Jingwen, 2018. Information extraction and knowledge graph construction from geoscience literature. Comput. Geosci. 112, 112–120. https://doi.org/10.1016/j.cageo.2017.12.007.
Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M., 2020. HuggingFace's Transformers: State-Of-The-Art Natural Language Processing arXiv:1910.03771 [cs].
Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V., 2020. XLNet: Generalized Autoregressive Pretraining for Language Understanding arXiv:1906.08237 [cs].
Zhang, D., Yuan, Z., Liu, Y., Zhuang, F., Chen, H., Xiong, H., 2021. E-BERT: A Phrase and Product Knowledge Enhanced Language Model for E-Commerce arXiv:2009.02835 [cs].
Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S., 2015. Aligning Books and Movies: towards Story-like Visual Explanations by Watching Movies and Reading Books arXiv:1506.06724 [cs].