SOIL Discuss., https://doi.org/10.5194/soil-2018-44
Manuscript under review for journal SOIL
Discussion started: 29 January 2019

Word embeddings for application in geosciences: development, evaluation and examples of soil-related concepts

José Padarian and Ignacio Fuentes

Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, New South Wales, Australia

Correspondence: José Padarian (jose.padarian@syndey.edu.au)

Abstract. A large amount of descriptive information is available in most disciplines of geosciences. This information is usually considered subjective and ill-favoured compared with its numerical counterpart. Considering the advances in natural language processing and machine learning, it is possible to utilise descriptive information and encode it as dense vectors. These word embeddings lay on a multi-dimensional space where angles and distances have a linguistic interpretation. We used 280,764 full-text scientific articles related to geosciences to train a domain-specific language model capable of generating such embeddings. To evaluate the quality of the numerical representations, we performed three intrinsic evaluations, namely: the capacity to generate analogies, term relatedness compared with the opinion of a human subject, and categorisation of different groups of words. Since this is the first attempt to evaluate word embedding for tasks in the geosciences domain, we created a test suite specific for geosciences. We compared our results with general domain embeddings commonly used in other disciplines. As expected, our domain-specific embeddings (GeoVec) outperformed general domain embeddings in all tasks, with an overall performance improvement of 107.9%. The resulting embedding and test suite will be made available for other researchers to use an expand.

1 Introduction

Whilst different machine learning methods have been used in geosciences (Lary et al., 2016), natural language processing (NLP) techniques, which involve the manipulation and analysis of language (Jain et al., 2018), have rarely been applied. This is mainly due to the prioritisation of numerical data over qualitative descriptions, which are usually considered of subjective nature (McBratney and Odeh, 1997). However, it must be taken into account the resources that have been invested in collecting large amounts of descriptive information from pedological, geological and other fields of geosciences. Neglecting non-numerical data due to its bias or inconsistency seems wasteful. Moreover, considering the advances in NLP and machine learning, a significant fraction of the subjectivity and ambiguity introduced by language can be removed by text processing and probabilistic analysis (Bakx et al., 2006).

For soil sciences, the use of NLP opens the possibility to a broad range of new analyses. Some examples include general, discipline-wide methods such as automated content analysis (Nunez-Mir et al., 2016) or recommendation systems (Wang and Blei, 2011) which can take advantage of the current literature. More specific cases could take advantage of big archives of descriptive data, like the ones reported by Arrouays et al. (2017). The authors mention examples such as the Netherlands with more than 327,000 auger descriptions covering agricultural, forest and natural lands, or north-central US with 47,364 pedon descriptions covering 8 states.

Approaches to deal with descriptive data include the work of Fonseca et al. (2002) who proposed the use of ontologies to integrate geographic information of different kinds. At the University of Colorado, Chris Jenkins created a structured vocabulary for geomaterials (http://instaar.colorado.edu/~jenkinsc/dbseabed/resources/geomaterials/) using lexical extraction (Miller, 1995), names decomposition (Peckham, 2014) and distributional semantics (Baroni et al., 2012) in order to characterise word terms for use in Natural Language Processing and other applications. A different approach, perhaps closer to the preferred quantitative methods, is the use of dense word embeddings (vectors) which encode information about a word and its linguistic relationships with other words, positioning it on a multi-dimensional space. The latter is the focus of this study.

There are many general-purpose word embeddings trained on large corpora from social media or knowledge organisation archives such as Wikipedia (Pennington et al., 2014; Bojanowski et al., 2016). These embeddings have been proven to be useful in many tasks such as machine translation (Mikolov et al., 2013a), video description (Venugopalan et al., 2016), document summarisation (Goldstein et al., 2000), and spell checking (Pande, 2017). However, for field-specific tasks, word embeddings trained on specialised corpora can capture the semantics of terms better than those trained on general corpora (Pakhomov et al., 2016; Wang et al., 2018).

As far as we are aware, this is the first attempt to develop and evaluate word embedding for the geosciences domain. This paper is structured as follow: first, we define what word embeddings are, explaining how they work and showing examples to help the reader understand some of their properties. Second, we describe the text data used and the pre-processes required to train a language model and generate these word embedding (GeoVec). Third, we illustrate how a natural language model can be quantitatively evaluated and we present the first test dataset for the evaluation of word embeddings specifically developed for the geosciences domain. Fourth, we present result of an intrinsic evaluation of our language model using our test dataset. Finally, we explore some of the characteristics of the multi-dimensional space and the linguistic relationships captured by the model through examples of soil-related concepts.

2 Word embeddings

Word embedding have been commonly used in many science disciplines, thanks to their application in statistics. For example, one-hot encodings (Fig. 1), also know as “dummy variables”, have been used in regression analysis since at least 1957 (Suits, 1957). In one-hot encoding, each word is represented by a vector of length equal to the number of classes or words, where each dimension represents a feature. The problem with this representation is that the resulting array is sparse (mostly zeros) and very large when using a large corpora, and also presents problem of poor estimation of the parameters of the less-common words (Turian et al., 2010). A solution for these problems is the use of unsupervised learning to induce dense, low-dimensional embeddings (Bengio, 2008). The resulting embeddings lay on a multi-dimensional space where angles and distances have a linguistic interpretation. One of the most common examples of this property is the capacity of generating analogies with vector arithmetic. For instance, the analogy "king is to queen as man is to woman" could be represented by the vector equality king − queen = man − woman.

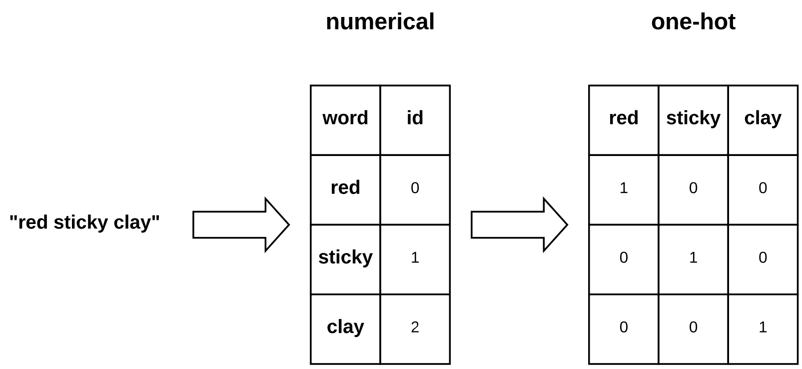

Figure 1. Example of two encodings of the phrase "red sticky clay", numerical and one-hot.

This and other relationships can be extracted from the embeddings where, potentially, each dimension and interaction within the high-dimensional space encodes a different type of relationship. A more complex example is the representation of the country-capital relationship. Fig. 2 presents a principal component analysis (PCA) projection of pairs of words with such relationship. Without explicitly enforcing this relationship when creating the language model, the resulting word embeddings encode the country-capital relationship due to the high co-occurrence of the terms. In Fig. 2 is also possible to observe a second relationship, geographic location, where South American countries are positioned to the right, European countries in the middle and (Eur-)Asian countries to the left.

3 Data, text pre-processing and model training

3.1 Corpus

The corpus was generated by retrieving and processing 280,764 full-text articles related to geosciences. We used the Elsevier ScienceDirect APIs to search for manuscript that matched the terms listed in Table 1. We also included Wikipedia articles which list and concisely define some concepts like types of rocks, minerals, and soils. We downloaded the text from Wikipidia articles "List_of_rock_types", "List_of_minerals", "List_of_landforms", "Rock_(geology)", "USDA_soil_taxonomy" and "FAO_soil_classification", and also all the Wikipedia articles linked from those pages.

SOIL Discuss., https://doi.org/10.5194/soil-2018-44
Manuscript under review for journal SOIL
Discussion started: 29 January 2019

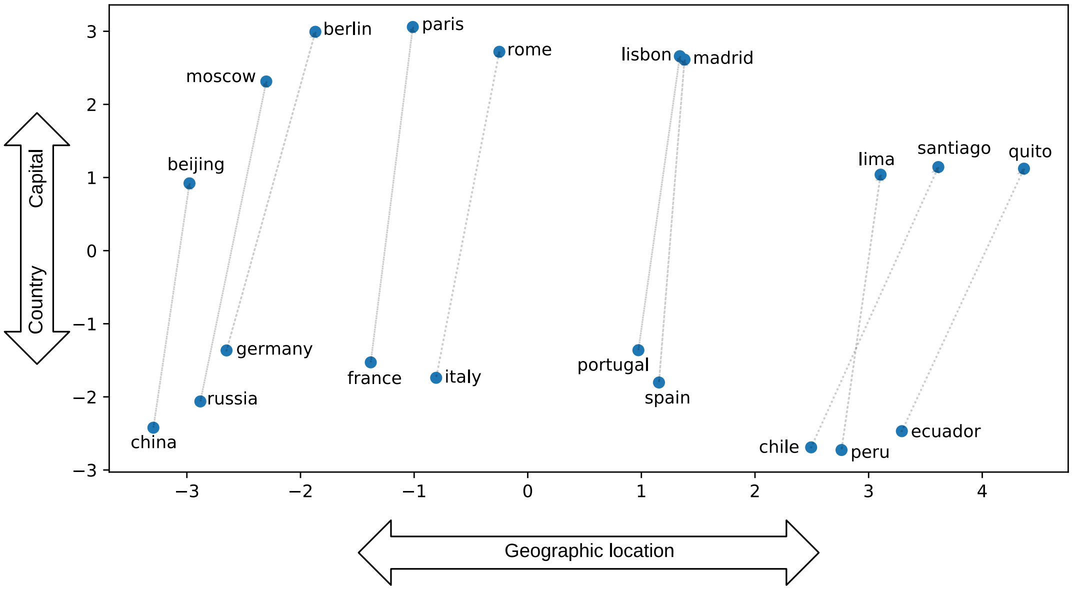

Figure 2. Examples of two-dimensional PCA projection of selected word embeddings using a general domain model. The figure illustrates the country-capital relationship learned by the model. Also notice that the model learned about the geographic relationship between the places. Example adapted from Mikolov et al. (2013b).

3.2 Pre-processing

The corpus was split in sentences which were then pre-processed using a sequence of commonly used procedures including: *a)* removing punctuation, *b)* lower-casing, *c)* removing digits and symbols, and *d)* removing (easily identifiable) references. The clean sentences were then tokenised (split into words) stemmed using Porter's algorithm (Porter, 1980) in order to remove some derivational affixes (e.g.: coating → coat). We excluded words with less than 3 characters, and we removed common English words such as 'not', 'only' , and 'most' (a full list can be found in the documentation of the nltk python library (Bird and Loper, 2004)). Finally, we excluded sentences with less than 3 words. The final corpus has a vocabulary size of 701,415 (unique) words and 305,290,867 tokens.

3.3 Model training

For this work we used the GloVe (Global Vectors) model (Pennington et al., 2014), developed by Stanford University NLP group, which achieved great accuracy on word analogy tasks and outperformed other word embedding models on similarity and entity recognition tasks. As many NLP methods, GloVe relays on ratios of word-word co-occurrence probabilities in the corpus, generating a vector space with the linear substructures mentioned in Section 2. To calculate the co-occurrence probabilities, GloVe uses a local context window, where a pair of words d words apart contributes to a $1/d$ to the total count.

SOIL Discuss., https://doi.org/10.5194/soil-2018-44
Manuscript under review for journal SOIL
Discussion started: 29 January 2019

Table 1. Search terms used to retrieve full-text articles from Elsevier ScienceDirect APIs.

Search terms		
Acrisol	Geosciences	Permafrost
Alfisol	Groundwater	Petrology
Allophane	Gypsisols	Podzols
Andisol	Histosol	Sedimentary
Andosols	Hydrogeology	Sedimentary mineralogy
Aridisol	Igneous petrology	Sedimentary petrology
Chernozems	Imogolite	Sedimentary rocks
Entisol	Inceptisol	Sedimentology
Environmental geology	Lithology	Soil classification
Field geology	Metamorphic petrology	Spodosol
Gelisol	Mineralogy	Stratigraphy
Geochemistry	Mollisol	Ultisol
Geology	Oxisol	Vertisol
Geomaterials	Peatland	Volcanic soils
Geomorphology	Pedogenesis	
Geophysics	Pedology	

We trained the model during 60 epochs, where 1 epoch corresponds to a complete pass through the training dataset. During the training phase we experimented using embedding of different number of components (dimensions) and different context window sizes. Here we present the results for 300 components and a window size of 10, which represents a good balance between model size, training time and performance.

4 Evaluation of word embeddings

Given the characteristic of the vector space, the most common method to evaluate word embeddings is to asses their performance in tasks that test if semantic and syntactic rules are properly encoded. Many studies have presented datasets to perform this task. Rubenstein and Goodenough (1965) presented a set of 65 noun synonyms to test the relationship between the semantic similarity existing between a pair of words and the degree to which their contexts are similar. More recent and larger test dataset and task types have been proposed (Finkelstein et al., 2002; Mikolov et al., 2013c; Baroni et al., 2014) but they all have been designed with the aim to test general domain vectors. Because the aim of this work is to generate embeddings for the geosciences domain, we developed a test suite to evaluate their intrinsic quality in different tasks, which are described bellow.

Analogy: Given two related pairs of words, a:b and x:y, the aim of the task is to answer the question "a is to x as b is to?".

The set includes 50 quartets of words with different levels of complexity, from simple semantic relationships to more advance syntactic relations. In practice, is possible to find y by calculating the cosine similarity between the differences of the paired vectors:

$$\frac{(v_b - v_a) \cdot (v_y - v_x)}{\|v_b - v_a\| \|v_y - v_x\|} \tag{1}$$

In this case v_y is the embedding for each word of the vocabulary and y is the word with the highest cosine similarity. Some examples of analogies are: “moraine is to glacial as terrace is to ___? (fluvial)”, “limestone is to sedimentary as tuff is to ___? (volcanic)” and “chalcantite is to blue as malachite is to ___? (green)”.

We estimated the top-1, top-3, top-5 and top-10 accuracy score, recording a positive result if y was within the first 1, 3, 5 or 10 words returned by the model, respectively.

Relatedness: For a given pair of words (a,b), a score of 0 or 1 is assigned by a human subject if the words are unrelated or related, respectively. The set includes 100 pairs of scored pairs of words. The scores are expected to have a high correlation with the cosine similarity between the embeddings of each pair of words. In this work we used the Pearson correlation coefficient.

Categorisation: Given 2 sets of words $s_1 = \{a,b,c,...\}$ and $s_2 = \{x,y,z,...\}$, this test should be able to correctly assign each word to its corresponding group using a clustering algorithm. We provide 30 tests with 2 clusters each. We estimated the v-measure score (Rosenberg and Hirschberg, 2007), which takes into account the homogeneity and completeness of the clusters, after projecting the multi-dimensional vector space to a two-dimensional PCA space and performing a k-means clustering. Given that k-means is not deterministic (when using random centroids initiation), we used the mean v-measure score of 50 realisations.

We compared our results with general domain vectors trained on Wikipedia articles (until 2014) and the Gigaword v5 catalogue, which comprise 6 billion tokens and is provided by the authors of GloVe at https://nlp.stanford.edu/projects/glove/.

5 Results and discussion

5.1 Co-occurrence

Before training the language model, the first output of the process is a co-occurrence matrix. This matrix encodes useful information about the underlying corpus (Heimerl and Gleicher, 2018). Fig. 3 shows the co-occurrence probabilities of soil taxonomic orders and some selected words. Is possible to observe that concepts generally associated with a specific order actually co-occur in the corpus, such as soil cracks, which are features usually present in Vertisols; or Andisols being closely related to areas with volcanic activity.

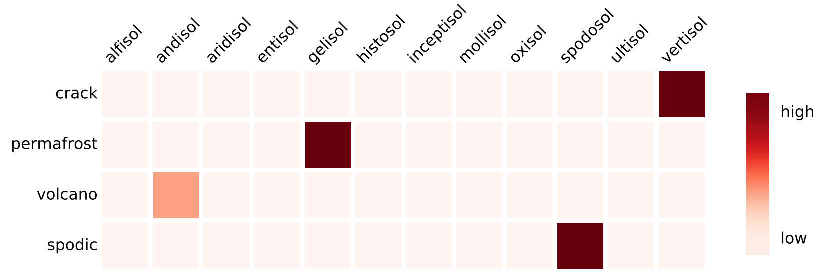

Figure 3. Co-occurrence probability matrix of soil orders (USDA) and selected words.

This information can also be used to guide the process of generating a domain-specific model. In our case, in an early stage of this study, the terms “permafrost” and “gelisol” presented a very low co-occurrence probability, a clear sign of the limited topic coverage of the articles at that point.

5.2 Intrinsic evaluation

The results of the intrinsic evaluation indicate that our domain-specific embeddings (GeoVec) performed better than the general domain embeddings in all tasks (Table 2), increasing the overall performance by 107.9%. This is an expected outcome considering the specificity of the tasks. For the analogies, we decided to present the top-1, 3, 5 and 10 accuracy scores because, even if the most desirable result is to have the expected word as the first output from the model, in many cases the first few words are closely related or they are synonyms. For instance, for the analogy “fan is to fluvial as estuary is to ____? (coastal)”, the first four alternatives are “tidal”, “river”, “estuarine”, “coastal”, which are all related to a estuary.

Table 2. Evaluation scores for each task for our domain-specific (GeoVec) and general domain embeddings (Stanford). For the analogy task, top-1, 3, 5 and 10 represents the accuracy if the expected word was within the first 1, 3, 5 or 10 words returned by the model. For the relatedness task, the score represents the absolute value of the Pearson correlation. For the categorisation task, the score represents the mean value of 50 v-measure scores. The possible range of all scores is 0 to 1, where higher is better.

	GeoVec	Stanford
Analogy (top-1)	0.39	0.22
Analogy (top-3)	0.78	0.37
Analogy (top-5)	0.90	0.41
Analogy (top-10)	0.92	0.49
Relatedness	0.61	0.23
Categorisation	0.75	0.38
Overall	0.73	0.35

SOIL Discuss., https://doi.org/10.5194/soil-2018-44
Manuscript under review for journal SOIL
Discussion started: 29 January 2019

It was possible to observe an increase on the overall performance of the embeddings (calculated as the mean of the analogy (top-5), relatedness and categorisation tasks) as we added more articles, almost stabilising around 300 million tokens, specially for the analogy task (Fig. 4). For domain-specific embeddings, this limit most likely varies depending on the task and domain. For instance, Pedersen et al. (2007), measuring semantic similarity and relatedness in the biomedical domain, found a limit of around 66 million tokens.

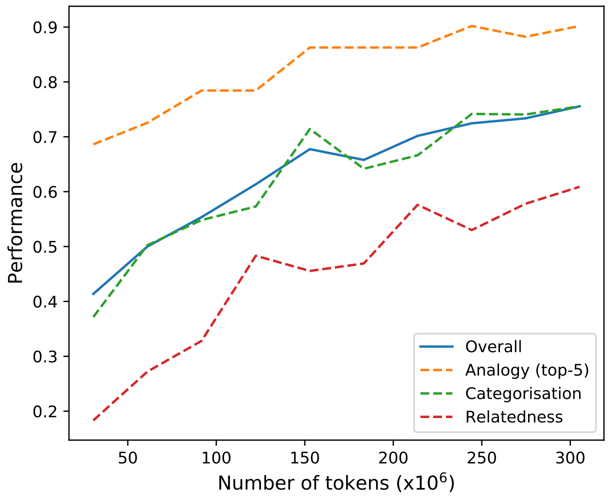

Figure 4. Overall performance of the embeddings versus number of tokens used to construct the co-occurrence matrix. The improvement limit is around 300 million tokens. For future comparisons, this limit corresponds to approximately: 280,000 articles, 22.5 million sentences and 700,000 unique tokens.

The improvement over the general domain embeddings has also been reported in other studies. Wang et al. (2018) concluded that word embeddings trained on biomedical corpora can capture the semantics of medical terms better than the embeddings of a general domain GloVe model. Also in a biomedical application, Jiang et al. (2015) and Pakhomov et al. (2016) reported similar conclusions. In the following sections we explore the characteristics of the obtained embeddings, showing some graphical examples of selected evaluation tasks.

5.3 Analogy

A different way of evaluating analogies is to plot the different pairs of words in a 2-dimensional PCA projection. Fig. 5 shows different pairs of words which can be seen as group analogies. From the plot, any pair of related words can be expressed as an analogy. For example, from the left panel, is possible to generate the analogy "claystone is to clay as sandstone is to ____? (sand)" and the first model output is indeed "sand".

As we showed in Fig. 2, the embeddings encode different relationships with different degrees of sophistication. In the left panel of Fig. 5 is possible to observe simple analogies, mostly syntactics since "claystone" contains the word "clay". The right panel presents a more advanced relationship where rock names are assigned to their corresponding rock type.

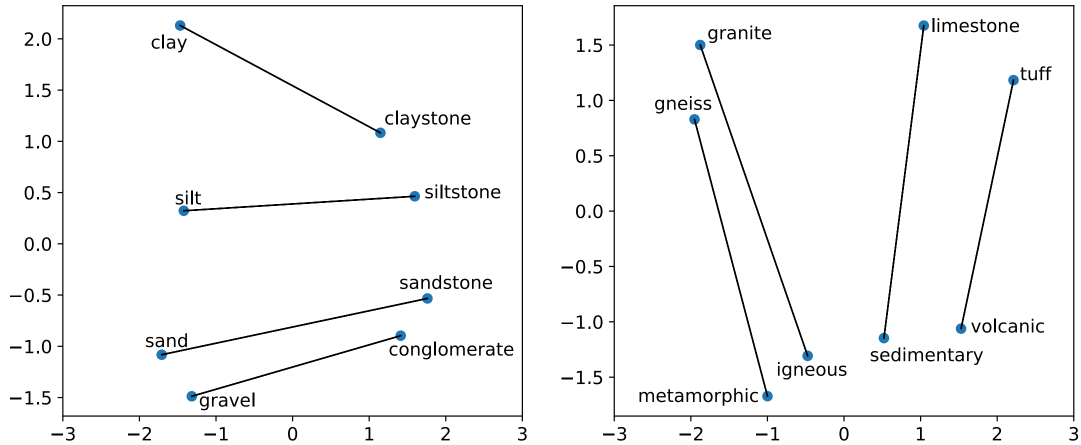

Figure 5. Two-dimensional PCA projection of selected words. Simple syntactic relationship between particle fraction sizes and rocks (left panel) and advanced semantic relationship between rocks and rock types (right panel).

5.4 Categorisation

Similar to the analogies, the categorisation task can also present different degrees of complexity of the representations. In the left panel of Fig. 6, a k-means clustering can distinguish the two expected clusters of concepts, WRB (FAO, 1988) and Soil Taxonomy (USDA, 2010) soil classification names. Andisols and Andosols are correctly assigned to their corresponding groups but apart from the rest, probably due to their unique characteristics. Vertisols are correctly placed in between the two groups, since both have a soil type with that name. A second level of aggregation can be observed in the right panel. The k-means clustering correctly assigned the same soil groups from the left panel into a general “soil types” group, different from “rocks”.

5.5 Other embedding properties

Interpolation of embeddings is an interesting exercise that has been used to generate gradient between faces (Yeh et al., 2016; Upchurch et al., 2017), assist drawing (Baxter and ichi Anjyo, 2006) and transform speech (Hsu et al., 2017). Interpolation between text embeddings are less common. Bowman et al. (2015) analysed the latent vector space of sentences and found that their model was able to generate coherent and diverse sentences when sampling between two embeddings. Duong et al. (2016) interpolated between embedding from two vector spaces trained on different languages corpora to create a single cross-lingual vector space. The vector space from our model also presents similar characteristics.

We were able to interpolate between different words, obtaining coherent concepts (Fig. 7). The interpolation between “clay” and “’boulder’, with fine and coarse size, respectively, yields a gradient of sizes, with “clay”<“silt”<“sand”<“gravel”<“cobble”<“boulder”. Another interpolation example, along another type of relationship, is shown in the right panel of Fig. 7. The interpola-

SOIL Discuss., https://doi.org/10.5194/soil-2018-44
Manuscript under review for journal SOIL
Discussion started: 29 January 2019

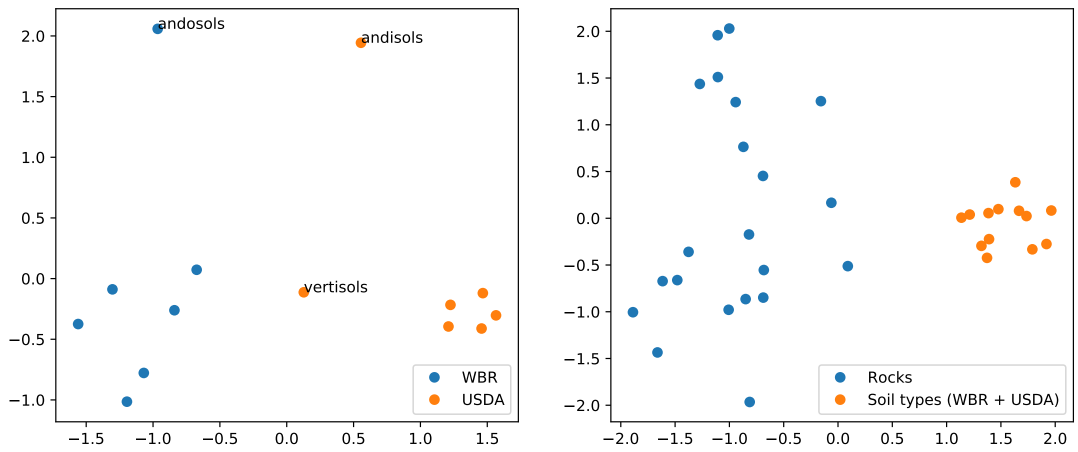

Figure 6. Two-dimensional PCA projection of selected categorisations. Clusters representing soil types from different soil classification systems (left panel) and a different aggregation level where the same soil types are grouped as a single cluster when compared with rocks (right panel).

tion between the rocks "slate" and "migmatite" yields a gradient of rocks with different grades of metamorphism, with "slate"<"phyllite"<"schist"<"gneiss"<"migmatite".

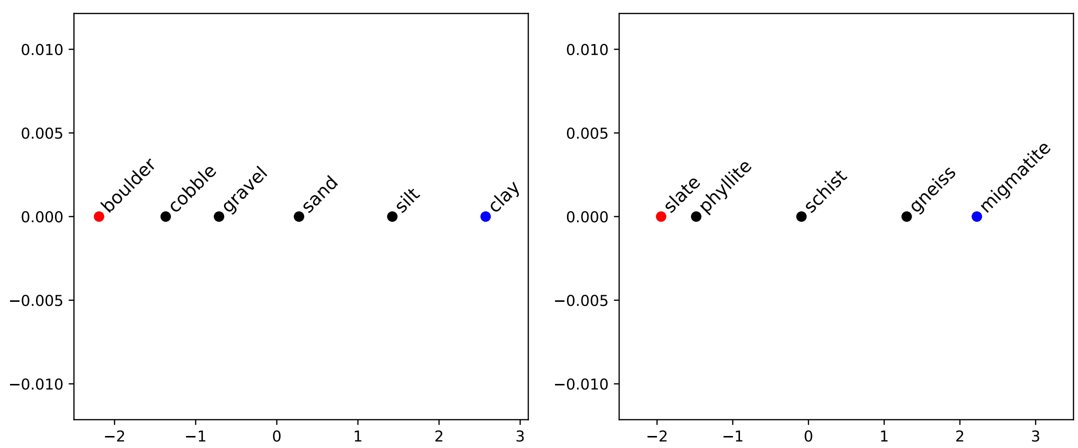

Figure 7. Interpolated embedding in a two-dimensional PCA projection showing a size gradient (left panel) with "clay"<"silt"<"sand"<"gravel"<"cobble"<"boulder"; and gradient of metamorphism grade (right panel) with "slate"<"phyllite"<"schist"<"gneiss"<"migmatite".

5.6 Future work

In the future, we expect to evaluate the effect of using our embeddings in downstream applications (extrinsic evaluation). It is expected that domain-specific embedding will necessarily improve the results of downstream tasks but this is not always the case. Schnabel et al. (2015) suggested that extrinsic evaluation should not be used as a proxy for a general notion of embedding quality, since different tasks favour different embeddings, but they are useful in characterising the relative strengths of different models. We also expect to expand the test suite with more diverse and complex tests, opening the process to the scientific community. Another interesting opportunity is the inclusion of word embeddings in numerical classification systems (Bidwell and Hole, 1964; Crommelin and De Gruijter, 1973; Sneath et al., 1973; Webster et al., 1977; Hughes et al., 2014) which try to remove subjectivity by classifying an entity (soil, rock, etc.) based on numerical attributes that describe its composition.

6 Conclusions

In this work we introduced the use of domain-specific word embeddings for geosciences (GeoVec) as a way to a) reduce subjectivity of descriptive data, and b) open the alternative to include such data into numerical data analysis. Comparing the result with general domain embeddings, trained on corpus such as Wikipedia, the domain-specific embedding performed better in common natural language processing tasks such as analogies, terms relatedness and categorisation, improving the overall accuracy by 107.9%.

We also presented a test suite, specifically designed for geosciences, to evaluate embedding intrinsic performance. It comprises tests for three tasks (analogy, relatedness and categorisation) with different levels of complexity. We expect to expand the test suite with more diverse and complex tests, opening the process to the scientific community.

We demonstrated that the high-dimensional space generated by the language model encodes different type of relationships, through examples of soil-related concepts. These relationships can, potentially, be used in novel downstream applications usually reserved for numerical data. Beyond the analytical opportunities provided by word embeddings, they are also an interesting way of exploring how a scientific community creates its own language and the interactions between domain-specific concepts.

Code availability. The embeddings, test suite and helper functions will be available at https://github.com/spadarian/GeoVec

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This research was supported by Sydney Informatics Hub, funded by the University of Sydney.

References

Arrouays, D., Leenaars, J. G., de Forges, A. C. R., Adhikari, K., Ballabio, C., Greve, M., Grundy, M., Guerrero, E., Hempel, J., Hengl, T., et al.: Soil legacy data rescue via GlobalSoilMap and other international and national initiatives, GeoResJ, 14, 1–19, 2017.

Bakx, G. E., Villodre, L., and Claramunt, G.: Machine learning techniques for word sense disambiguation, Unpublished doctoral dissertation, Universitat Politecnica de Catalunya, 2006.

Baroni, M., Bernardi, R., Do, N.-Q., and chieh Shan, C.: Entailment above the word level in distributional semantics, in: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 23–32, Association for Computational Linguistics, 2012.

Baroni, M., Dinu, G., and Kruszewski, G.: 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), vol. 1, pp. 238–247, 2014.

Baxter, W. and ichi Anjyo, K.: Latent doodle space, in: Computer Graphics Forum, vol. 25, pp. 477–485, Wiley Online Library, 2006.

Bengio, Y.: Neural net language models, Scholarpedia, 3, 3881, 2008.

Bidwell, O. and Hole, F.: Numerical taxonomy and soil classification, Soil Science, 97, 58–62, 1964.

Bird, S. and Loper, E.: NLTK: the natural language toolkit, in: Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, p. 31, Association for Computational Linguistics, 2004.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T.: Enriching Word Vectors with Subword Information, arXiv preprint arXiv:1607.04606, 2016.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., and Bengio, S.: Generating sentences from a continuous space, arXiv preprint arXiv:1511.06349, 2015.

Crommelin, R. D. and De Gruijter, J.: Cluster analysis applied to mineralogical data from the coversand formation in the Netherlands, Tech. rep., Stichting voor Bodemkartering Wageningen, 1973.

Duong, L., Kanayama, H., Ma, T., Bird, S., and Cohn, T.: Learning crosslingual word embeddings without bilingual corpora, arXiv preprint arXiv:1606.09403, 2016.

FAO: FAO/UNESCO Soil Map of the World. Revised legend, with corrections and updates. World Soil Resources Report 60, 1988.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E.: Placing search in context: The concept revisited, ACM Transactions on information systems, 20, 116–131, 2002.

Fonseca, F. T., Egenhofer, M. J., Agouris, P., and Câmara, G.: Using ontologies for integrated geographic information systems, Transactions in GIS, 6, 231–257, 2002.

Goldstein, J., Mittal, V., Carbonell, J., and Kantrowitz, M.: Multi-document summarization by sentence extraction, in: Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization, pp. 40–48, Association for Computational Linguistics, 2000.

Heimerl, F. and Gleicher, M.: Interactive analysis of word vector embeddings, in: Computer Graphics Forum, vol. 37, pp. 253–265, Wiley Online Library, 2018.

Hsu, W.-N., Zhang, Y., and Glass, J.: Learning latent representations for speech generation and transformation, arXiv preprint arXiv:1704.04222, 2017.

Hughes, P. A., McBratney, A. B., Minasny, B., and Campbell, S.: End members, end points and extragrades in numerical soil classification, Geoderma, 226, 365–375, 2014.

Jain, A., Kulkarni, G., and Shah, V.: Natural language processing, International Journal of Computer Sciences and Engineering, 6, 2018.

Jiang, Z., Li, L., Huang, D., and Jin, L.: Training word embeddings for deep learning in biomedical text mining tasks, in: Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on, pp. 625–628, IEEE, 2015.

Lary, D. J., Alavi, A. H., Gandomi, A. H., and Walker, A. L.: Machine learning in geosciences and remote sensing, Geoscience Frontiers, 7, 3–10, 2016.

McBratney, A. B. and Odeh, I. O.: Application of fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions, Geoderma, 77, 85–113, 1997.

Mikolov, T., Le, Q. V., and Sutskever, I.: Exploiting similarities among languages for machine translation, arXiv preprint arXiv:1309.4168, 2013a.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J.: Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, pp. 3111–3119, 2013b.

Mikolov, T., tau Yih, W., and Zweig, G.: Linguistic regularities in continuous space word representations, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 746–751, 2013c.

Miller, G. A.: WordNet: a lexical database for English, Communications of the ACM, 38, 39–41, 1995.

Nunez-Mir, G. C., Iannone, B. V., Pijanowski, B. C., Kong, N., and Fei, S.: Automated content analysis: addressing the big literature challenge in ecology and evolution, Methods in Ecology and Evolution, 7, 1262–1272, 2016.

Pakhomov, S. V., Finley, G., McEwan, R., Wang, Y., and Melton, G. B.: Corpus domain effects on distributional semantic modeling of medical terms, Bioinformatics, 32, 3635–3644, 2016.

Pande, H.: Effective search space reduction for spell correction using character neural embeddings, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, vol. 2, pp. 170–174, 2017.

Peckham, S. D.: The CSDMS standard names: Cross-domain naming conventions for describing process models, data sets and their associated variables, 2014.

Pedersen, T., Pakhomov, S. V., Patwardhan, S., and Chute, C. G.: Measures of semantic similarity and relatedness in the biomedical domain, Journal of biomedical informatics, 40, 288–299, 2007.

Pennington, J., Socher, R., and Manning, C.: Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543, 2014.

Porter, M. F.: An algorithm for suffix stripping, Program, 14, 130–137, 1980.

Rosenberg, A. and Hirschberg, J.: V-measure: A conditional entropy-based external cluster evaluation measure, in: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), 2007.

Rubenstein, H. and Goodenough, J. B.: Contextual correlates of synonymy, Communications of the ACM, 8, 627–633, 1965.

Schnabel, T., Labutov, I., Mimno, D., and Joachims, T.: Evaluation methods for unsupervised word embeddings, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 298–307, 2015.

Sneath, P. H., Sokal, R. R., et al.: Numerical taxonomy. The principles and practice of numerical classification., 1973.

Suits, D. B.: Use of dummy variables in regression equations, Journal of the American Statistical Association, 52, 548–551, 1957.

Turian, J., Ratinov, L., and Bengio, Y.: Word representations: a simple and general method for semi-supervised learning, in: Proceedings of the 48th annual meeting of the association for computational linguistics, pp. 384–394, Association for Computational Linguistics, 2010.

Upchurch, P., Gardner, J. R., Pleiss, G., Pless, R., Snavely, N., Bala, K., and Weinberger, K. Q.: Deep Feature Interpolation for Image Content Changes., in: CVPR, vol. 1, p. 3, 2017.

USDA, N.: Keys to soil taxonomy, Soil Survey Staff, Washington, 2010.

Venugopalan, S., Hendricks, L. A., Mooney, R., and Saenko, K.: Improving LSTM-based video description with linguistic knowledge mined from text, arXiv preprint arXiv:1604.01729, 2016.

Wang, C. and Blei, D. M.: Collaborative topic modeling for recommending scientific articles, in: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 448–456, ACM, 2011.

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

Webster, R. et al.: Quantitative and numerical methods in soil classification and survey., Quantitative and Numerical Methods in Soil Classification and Survey., 1977.

Yeh, R., Liu, Z., Goldman, D. B., and Agarwala, A.: Semantic facial expression editing using autoencoded flow, arXiv preprint arXiv:1611.09961, 2016.