SOIL Discuss.,
https://doi.org/10.5194/soil-2018-44-RC1, 2019

Interactive comment on "Word embeddings for application in geosciences: development, evaluation and examples of soil-related concepts" *by* José Padarian and Ignacio Fuentes

Diana Maynard (Referee)

d.maynard@sheffield.ac.uk

Received and published: 9 February 2019

In this paper, the authors train some standard word embeddings specifically on a geoscience corpus and show that, unsurprisingly, these are better than some pre-existing embeddings trained on a general corpus. This is well-known, so there are no particularly interesting findings specifically in that result. Making the embeddings available to others could be beneficial, but for this a larger set would ideally be required, and certainly more proof of their usefulness would be needed.

While the application of word embeddings-based language analysis techniques is relatively new in the field of geo-sciences, the authors do not provide sufficient scientific

contribution or motivation in this paper. I have two major issues with the work presented: 1. It's not clear what they want to use word embeddings specifically for. They experiment with training some existing techniques on a geoscience corpus, but there is no actual motivation for doing so. Word embeddings are only useful if they are applied to a specific task, and if it can be shown that they help to solve the task in a better way than existing techniques. But the authors give no real-world task and just evaluate the quality of the embeddings on standard fun tasks such as analogies that have no actual purpose. Creating a good set of embeddings is one thing, but half the task lies in finding the best way to transform the topic vectors. 2. The authors do not do anything that involves scientific novelty - they simply take some existing word embeddings models and train them on a new corpus, which requires no novel critical thinking. This is useful therefore only as a means to an end, but is not worthy of publication in itself. The work could be interesting if it were taken a step further, but currently it is insufficient.

Some more specific points follow.

The abstract does not make it clear what the motivation for the work is, beyond the fact that word embeddings have not been trained on such a corpus, but this is insufficient justification.

The introduction is vague, e.g. "different machine learning methods have been used for geoscience" - what does this tell us? Nothing. We need to know at least what task they have been used for, why they have been used, and how well they work, not to mention why it is relevant to the work presented. Similarly, much of it is too imprecise, e.g. "subjectivity and ambiguity introduced by language can be removed by text processing and probabilistic analysis" - this needs clarification. Subjectivity can almost never be "removed" by NLP techniques, and as for ambiguity, this depends a lot on the kind of ambiguity and the task - and typically only those things which are ambiguous to computers but not humans can be dealt with successfully.

Again, the introduction should explain the motivation behind the work, why word embeddings are useful, and give some idea of the kinds of tasks they are going to be used for here. References to related work are lacking - the authors need to do proper research into the state of the art here - for example, properly investigating the advantages/disadvantages of training word embeddings on a general vs specific corpus.

The section on word embeddings is neither a clear general explanation for those who have no idea what they are (as one might expect in the geoscience field), nor does it provide a technical explanation for those familiar with the topic. The authors introduce the idea of analogies being produced with word embeddings, but do not explain why this is even interesting. Figures 1 and 2 are not clear to those who don't know already about word embeddings, and obsolete for those who do. In general, this section is very inadequate.

Section 3 is lacking in technical detail. How were the terms listed in Table 1 decided? Why were these particular pre-processing decisions taken? See for example (Denny and Spirling, 2018) on the importance of such decisions on the results obtained, and the effect that even small changes to these decisions can have on the end results. Denny, Matthew J., and Arthur Spirling. "Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it." Political Analysis 26.2 (2018): 168-189. For example, why do you use stemming and not morphological analysis? Surely you do not want to conflate tokens with different POS tags here? In other words, you want to perform inflectional but not derivational morphological analysis - this is more commonly used for pre-processing word embedding training than just stemming (the easy option). Either way, these decisions need to be properly justified.

Evaluation: You need to provide proper information here. For the relatedness task, who did the scoring? Was Inter-Annotator Agreement measured (and if not why not?)? What was the result of IAA? I would not expect high agreement here because this is a hard task for humans to perform, so this is really critical in order to have a valid set of gold standard data. How many tokens is your dataset also?

Section 5.5 - I suggest explaining what you mean by interpolation of embeddings.

The Conclusion section is very brief and, unsurprisingly given the rest of the paper, gives no real interesting conclusions. The final sentence is very unsatisfactory: what do you even mean by saying that embeddings give the scientific community an interesting way of "exploring how a scientific community creates its own language...."? You certainly haven't studied this in this work, and have no insights to offer us on it.

In summary, the development of a specific set of embeddings for geosciences could be useful, but this is all rather insufficient for publication here, and the relevance to geosciences, and specifically soil, is minimal. I suggest waiting until you have at least attempted to resolve some specific task using the embeddings, which relevant specifically to the topic of soil, before attempting to publish.

Interactive comment on SOIL Discuss., https://doi.org/10.5194/soil-2018-44, 2019.