SOIL Discuss.,
https://doi.org/10.5194/soil-2018-44-SC2, 2019

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

Alex McBratney

alex.mcbratney@sydney.edu.au

Received and published: 23 February 2019

First of all it is worth establishing the basis of the analysis. As I understand it - it seems to be based on the multidimensional scaling (principal coordinates) of a co-occurrence probability matrix - where the co-occurrences are related to a short list of pairs of words =(terms?) separated by a given word distance The words being meaningful in the soil science community. Is this a reasonable summary?

I guess one can ask what can we gain from this kind of analysis? It is very dependent on the words chosen. However it does show which words, and therefore concepts?, are used together. Does it reflect a pattern of usage? a similarity or difference of

concepts? or something more? From this can we quantify papers in terms of their content? Textual analysis is used to quantify free-form responses to survey questions. Scientific papers are of course less free form. I guess in the end my main question is - does this kind of analysis tell us anything about science or nature or does it really just tell us about the humanly constructed way that science is done and reported?

Looking at Figure 6. The diagram on the left looking at the names of soil 'orders' or 'reference groups' in a couple of systems shows that the two systems do not overlap - that the word 'vertisol' used in both systems split the difference - but of course the meaning/definition in the two systems might not be the same. It also suggests that volcanic soils are different in the two systems but are different from other soils. What do we learn - is there no overlap between the two systems? - if so then this is a complete disaster for soil classification. One interpretation is that it shows that there are two user communities and they do not cross-reference each other.Unhelpful for soil science and soil sustainability. The recent quantitative work in Geoderma by Hughes et al based on soil properties shows some degree of overlap between systems . It would be useful to label all the points on this diagram. The diagram on the right shows no overlap between soil classes and rock classes. It might suggest also that soils are more similar to each other than rocks, or that the way the words are used are more heterogeneous.

I could not quite understand Figure 7 - it shows meaningful continua of terms and in the correct order - is it a construction? or is it based on an analysis of papers? This reminds me of course of another approach - if one of the aims of the work here is to attempt to quantify meaning via words - then the fuzzy or continuous class approach is a good alternative, and perhaps should be compared. There are papers by the late Inakwu Odeh on this, and part a chapter in the recently-published Pedometrics book.

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