Review of “The Geometry of Culture: Analyzing Meaning through Word Embeddings”

Kozlowski et al (2018) have attempted a very interesting take at understanding culture through text. Word embeddings were developed as a tool to better understand *words* in a high dimensional space, as well as in performing vector operations on these projected words. While traditional work in word embeddings involved in using these word embeddings to carry out tasks such as learning the closest related entity, or to learn phrases, the *nature of the space* itself was not examined in as much detail.

Kozlowski et al (2018) propose a new method/computational approach using these word embedding models. If we could describe a particular research question which the article attempts to address, it would be - “Could one analyse cultural semantics, cultural change, and cultural differences using word embedding models”, or “How would one analyse cultural semantics, cultural change and cultural differences using word embedding models?”. Since the paper describes an approach and does not answer a particular question, it rather encourages one to explore the tool and the possibilities which the tool offer us. We will examine the claims made by the authors about the efficiency of the tool, and also whether these tools can be easily taken up sociologists and other social scientists to perform textual analysis.

We also note that since the paper is still in its review stage, there are some concerns about the structure of the paper - there are no figures placed in yet, with all the figures and tables at the end of the manuscript, and it also runs to 73 pages long with all the references. While the material is all relevant and provides a broad variety of examples to make the point that the tool has potential, choosing a subset of the examples would help in increasing the brevity and readability of the paper. We would suggest including a supplementary reading material or text with a portion of the examples and results so as to result in a shorter paper.

The introduction serves as nice way to ease the reader into the possibilities of using textual analysis to understand and explore culture, while maintaining that it has its limitations and that the “large data” nature of text has not been adequately harnessed, which is a valid assessment.

The section titled “Formal Text Analysis in the Study of Culture” begins by citing work done in interpreting text to form cultural understanding. The important ideas of semantic networks and topic models are introduced, which have long been ways to analyse unstructured text. This was also briefly introduced in the last section though, through Latent Semantic Analysis and Latent Semantic Indexing, and could maybe be combined, or one of the two could have been chosen. Again, while both the sets of papers are relevant, this is only in the interest of providing more brevity.

The next section moves on to describing the utility of word embedding models, as well as flesh out the existing work done with them while describing the structure of these models. Again, the context of where the article is published is very important here - while there does seem to be a lot of information in this section (and throughout the article so far), it could very well be condensed, suggesting the user to read the original word embeddings papers to be more mathematically familiar with the models. That being said, the presence of the more detailed explanations serves to be very useful for those uninitiated to the world of word embeddings. Examples such as the *king - man + woman = queen,* and the association of the word *gay* in different locations of the multi-dimensional space depending on the temporal context illustrate usefulness of word2vec models when understanding semantics. The mentions of how social networks can also be projected onto a high dimensional network is not substantiated with examples or given a lot of context; it does not seem to provide an increased understanding of *why* they are useful, and can be removed. The points made about *why* these kind of models are useful as compared to the older topic models and semantic networks are valuable as they are a very valid point - most of the older models work largely on word frequencies which might sometimes not be the best way to proceed as they don't often take word co-occurence (context!) into account.

The authors then move on their main idea - that using these word embeddings in different ways we can peek into the semantic structures of the textual corpus we've used. Their idea is quite ingenious in the way it builds on the idea of differences in these high dimensional vector spaces; in particular that there are certain dimensions which encapsulate certain social characteristics. The example of the gender dimension, best represented by the vector operations *"man - woman"*, or *"he - she" ,* capture semantic information of how gender is encoded in words in text. By taking an average of five to six such pairs which the authors believe best represent the dimension, they can then project values on this dimension and see where on this dimension do the words lie - if, after a cosine similarity check on a dimension we receive a positive value, it lies on the *masculine* area of the dimension, and *feminine* area of the dimension. The authors also note that these signs depend on whether it was "man - woman" or "woman - man" to create the dimensions.

It is in this section that the crux of their method is described, and there are many positive points to notice. Firstly, it scales to any kind of word embedding and isn't limited to only *word2vec* or *FastText* - allowing it to be used in a variety of contexts, with word embedding models trained using different techniques. It also introduces the cultural dimensions and the way they're created. The references used to strengthen their claims about the validity of their method sufficiently justify it; word embedding models have been shown to effectively capture biases implicit in text in the past.

The authors then go on to discuss why word embeddings are a relevant and important tool to be used in cultural analysis. taking ideas and theory from structuralism, relational field theory, intersectionality and practice oriented theories of language. Practice oriented theories of language suggest that words only attain meaning in context, or when used with other words. Since word embeddings are generally constructed by using words and noting their context by using a sliding window, this captures the idea quite well. It should be noted that the citations suggesting this might be a bit broad - the book/article by Austin (1962) is a rather broad reference and might not accurately represent all the ideas that the authors might want to push forward, though the nascent ideas do exist.

The multi-dimensional aspect of word embedding encapsulate some of the concepts of intersectionality and relational field theory as well. Many of Bourdieu's ideas are important here, and form the crux of the author's arguments and serve as an important theoretical base.

By taking various theoretical ideas from a variety of practices, they help strengthen their case without many of the conceptual pitfalls which might be incorporated if relying only on one entirely.

The data and methods section underlies the technical details of their method. Since the word embedding model largely depends on the data used to train it as well as the method used, it is important to spell this out. The word2vec model, GloVe model, and FastText are all employed, which are arguably three of the most cited and most used word embedding models developed. The datasets trained on include the Google Ngram dataset and subsets of it; the authors recognise the drawbacks and limitations of this dataset and address them, and the assertion that it still remains the most viable dataset is well backed-up. Word embedding models usually don't face the drawbacks which might normally arise by using this dataset, mainly because word embedding models rely more on context than on word frequencies.

The idea of cultural dimensions is also *relative to the dataset it is trained on.* This means we can measure how the dimensions *themselves* change when trained on different bodies of text. What remains possibly ambiguous is how well we can judge a body of text to represent that culture, but this is also a difficult question to answer. If we go ahead with the idea that the *size* of the corpus can compensate for the possibility that a corpus might not be the best representation of a culture, their usage of Google Ngrams seems more justified.

The following sections describe the results of the model in placing words on the dimension and compare it to results of a survey of humans asked to place the words on a scale. The possibility of misrepresenting the results may arise here, but the authors proceed with caution on what the results mean, and fairly appropriately address any discrepancies. For example, the subjects of the survey may not belong

to the same cultural group as the texts generated. The results are certainly very interesting, such as the location of various genres of music on these dimensions. The authors further reinforce the fact that these bodies of text still incorporate historical bias, but it so happens that these biases are also largely present in the results of the survey as well, as is illustrated with public perception of techno music being “white” music.

Again, we might want to pick a little bit more into the survey used to determine how well the word embeddings model perform. 398 subjects were used via the Mechanical Turk platform, and how well these subjects’ cultural outlooks represent millions of text documents may be questioned, but given the limitations of actually collecting more data on what “actual” social outlooks, it is difficult to suggest an alternative. The nature of the results are also significant enough to impress us to further explore the tool. The usage of checking the performance by classifying words based on the cultural dimension was also another strong point in their case.

Both the historical analysis as well as the cross cultural analysis between the US and UK all bring up very fascinating examples which point out the various ways these cultural dimensions can be used. Again, a possible complaint in this section could be the volume of examples used, which might be better served being in a visualisation. It is a lot of text to plough through which reinforces the same point and using a visual tool while moving the textual examples to supplementary material would help in increasing the brevity and readability of the paper. That being said - the examples themselves are very revealing, whether it is the movement of the word “engineer” across the gender dimension, or how “worker” or “commoner” differs in the class dimension between US and the UK.

The next section is a discussion section which sums up the technical and theoretical aspects as well as the results of their method.

The method itself is a very novel addition on the existing word embeddings literature and provides us with a whole new park to play around in. Because of the very *nature* of the method and the *nature* of word embeddings, it means we can mix and match this with a variety of texts from different cultures - whether it is across time or across space or across both. The authors only used English language texts in their analysis but the method can very easily be used across multiple languages if multi-linguals are working on the dataset, and also opens up an entire realm of socio-lingual analysis to understand how different languages represent words with similar semantics but possibly very different contexts, and can possibly provide insights about how society influences language and vice versa.

The purpose is not to prove any causality, but rather to survey or peer into how texts influence culture and how culture influences text - an almost Foucaultian analysis. For example, it may be possible to see who scientific texts different in various cultural contexts in the classical age and in the modern age - is it possible

that their might have indeed been a shift in *episteme?* Using cultural dimensions we might be able to possibly see this by analysing large bodies of scientific text. By serving as an exploratory tool with a strong theoretical and technical basis, Kozlowski et al have provided an important approach to the Computational Sociology literature. In this sense they have answered a research question of *what* can be done using word embeddings in a broader sense.

The possible drawbacks would be to question the nature of the data used, but these drawbacks have been addressed by the authors; it would always be better to have different kinds of data, more survey subjects, and use more data to create the word embedding model. This however leaves a lot of room for future scientists to work on, by collecting and identifying more antonym pairs for different cultural dimensions or by identifying more possible dimensions to analyse, or different spatial-temporal cross sections.