Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database

Edgar Altszyler*, Sidarta Ribeiro†, Mariano Sigman‡ and Diego Fernandez-Slezak *
*Laboratorio de Inteligencia Artificial Aplicada, Departamento de Computación, Universidad de uenos Aires - CONICET
†Instituto do Cérebro, Universidade Federal do Rio Grande do Norte, Natal, Brazil
‡Universidad Torcuato Di Tella - CONICET

***Abstract*—This summary presents the results obtained in our work, *Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database* [1].**

1. Background

The main idea behind word embeddings is that words with similar meanings tend to occur in similar contexts. Base in this hypothesis, *word embeddings* describe each word in a vectorial space, where words with similar meanings are located close to each other. Word embeddings have been extensively studied in large text datasets. However, only a few studies analyze semantic representations of small corpora, particularly relevant in single-person text production studies.

In our paper [1], we compare the two most used embeddings (Skip-gram and LSA) capabilities in this scenario, and we test both techniques to identify word associations in dream reports series.

1.1. Dream content analysis

Most of the newest dream content analysis methods are based on frequency word-counting of predefined categories in dreams reports [2]. A well known limitation of this approach is the impossibility of identifying the meaning of the counted words, which are determined by the context in which they appear. To tackle this problem, we set out to study the capabilities of word embeddings to capture relevant word associations in dream reports series. This is the first time in which word embeddings has been applied to dream content analysis.

2. Summary of Results

Firstly, we test LSA and skip-grams performance in two semantic task for different corpus size. As it is known that the optimal embeddings dimensions depends on the corpus size [3], we also vary the number dimensions and use the best result for each corpus size. We found that Skip-gram models has a steeper learning curve, outperforming LSA when the models are trained with medium to large datasets. However, when the corpus size is reduced, Skip-gram's performance has a severe decrease, thus LSA becoming the more suitable tool.

Secondly, we test word embeddings capabilities to identify word associations in dream reports series. In particular, we test whether these tools were able to capture accurately, in different manually annotated dream reports series, the semantic neighborhood of the word *run*. We found that LSA can effectively differentiate different word usage pattern even in cases of series with low number of dreams and low frequency of target words.

3. Conclusion

In our work, we show in two semantic tests that LSA is more appropriate in small-size corpus scenarios than the well-used skip-gram model. Also we show that LSA can accurately quantify words associations in dreams reports. This is a step forward in the application of word embeddings to the analysis of dream content. We propose that LSA can be used to explore word associations in dreams reports, which could bring new insight into this classic field of psychological research. On one hand, the validation of semantic metrics to analyze word associations in dream reports promises a much more accurate quantification of socially-shared meaning in dream reports, with great potential application in psychiatric diagnosis [4] and dream decoding research [5]

References

[1] Altszyler, E., Ribeiro, S., Sigman, M., and Fernandez-Slezak, D. Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database. Under review in consciousness and cognition journal (*arXiv preprint arXiv:1610.01520)*

[2] Domhoff, G. W. and Schneider, A (2008). Studying dream content using the archive and search engine on DreamBank.net.

[3] Fernandes, J., Artífice, A., and Fonseca, M. J. (2011). Automatic estimation of the lsa dimension.

[4] Mota, N. B., Furtado, R., Maia, P. P., Copelli, M., and Ribeiro, S. (2014). Graph analysis of dream reports is especially informative about psychosis.

[5] Horikawa, T., Tamaki, M., Miyawaki, Y., and Kamitani, Y. (2013). Neural decoding of visual imagery during sleep.