Consciousness and Cognition xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Consciousness and Cognition

journal homepage: www.elsevier.com/locate/concog

The interpretation of dream meaning: Resolving ambiguity using Latent Semantic Analysis in a small corpus of text

Edgar Altszyler[a,*], Sidarta Ribeiro[b], Mariano Sigman[c], Diego Fernández Slezak[a]

[a] Depto. de Computación, Universidad de Buenos Aires, Ciudad universitaria, CONICET, Pabellon 1, C1428EGA, Argentina
[b] Instituto do Cérebro, Universidade Federal do Rio Grande do Norte, Natal, Brazil
[c] Universidad Torcuato Di Tella – CONICET, Argentina

ARTICLE INFO

Keywords:
Dream content analysis
Word2vec
Latent Semantic Analysis

ABSTRACT

Computer-based dreams content analysis relies on word frequencies within predefined categories in order to identify different elements in text. As a complementary approach, we explored the capabilities and limitations of word-embedding techniques to identify word usage patterns among dream reports. These tools allow us to quantify words associations in text and to identify the meaning of target words. Word-embeddings have been extensively studied in large datasets, but only a few studies analyze semantic representations in small corpora. To fill this gap, we compared Skip-gram and Latent Semantic Analysis (LSA) capabilities to extract semantic associations from dream reports. LSA showed better performance than Skip-gram in small size corpora in two tests. Furthermore, LSA captured relevant word associations in dream collection, even in cases with low-frequency words or small numbers of dreams. Word associations in dreams reports can thus be quantified by LSA, which opens new avenues for dream interpretation and decoding.

1. Introduction

The analysis of dream contents marked the dawn of Psychology (Freud, 1900; Kraepelin, 1906). Dream contents show gender and cultural differences, consistency over time, and concordance with waking-life experiences, such as activity and emotions (Bell & Hall, 2011; Domhoff, 2002; Domhoff & Schneider, 2008a). Dream contents change after drug treatment (Kirschner, 1999) or due to psychiatric disorders (Domhoff, 2000). Along this line, recently, Mota, Furtado, Maia, Copelli, and Ribeiro (2014) have shown that the graph analysis of dreams reports is quite informative about psychosis, being useful to predict the Schizophrenia diagnosis (Mota, Copelli, & Ribeiro, 2017).

Dream content analysis has been employed to infer the mechanisms that shaped the evolution of dreaming. Threat Simulation Theory (TST) has been particularly influential. It describes the function of dreaming in terms of an evolutionarily selected mechanism, which provides a *world simulation* where we can train responses to threatening experiences (Revonsuo, 2000). This theory brings the analysis of dream contents centerstage, as a natural approach for the investigation of their functionality. For example, Valli et al. (2005) tested the Threat Simulation Theory by comparing the content of the dreams reported by traumatized and non-traumatized children. Their results show an increased number of threatening dream events in traumatized population, thus giving support to the Threat Simulation Theory. Despite the large empirical evidence supporting this theory, some contradictory evidence has been reported against TST (Malcolm-Smith, Koopowitz, Pantelis, & Solms, 2012; Malcolm-Smith, Solms, Turnbull, & Tredoux, 2008). In

* Corresponding author.
E-mail address: ealtszyler@dc.uba.ar (E. Altszyler).

http://dx.doi.org/10.1016/j.concog.2017.09.004
Received 6 March 2017; Received in revised form 25 August 2017; Accepted 10 September 2017

this sense, authors have suggested more structured and data-driven tests to be developed (Revonsuo, Tuominen, & Valli, 2015; Revonsuo & Valli, 2008; Valli, 2011). Thus, computational approaches could provide a quantitative framework to tests hypothesis derived from dreams theories.

Most of computational dream content analysis methods are based on frequency word-counting of predefined categories in dreams reports (Bulkeley, 2009; Domhoff & Schneider, 2008b). For example, these techniques have been successfully used to quantify the presence of emotions, sexual content and references to cognitive activity (Bulkeley, 2014). This approach is focused on the salience of words without identifying the context in which they appear. For instance, the occurrence of the word *fall* in a dream report may be used in different contexts, such as *falling* from a cliff, teeth *falling* out or *falling* sick. Moreover, since language is inherently polysemic (Sigman & Cecchi, 2002), semantic ambiguity due to polysemic word associations could hinder the automated analysis of dream reports. In this context, we set out to study the capabilities of word embeddings to capture relevant word associations in dream reports. We believe that word embeddings can be useful not only in extracting words meaning but also in establishing relationships between elements present in dreams.

Corpus-based semantic representations (i.e. embeddings) exploit statistical properties of textual structure to embed words in a vector space. In this space, terms with similar meanings tend to be located close to each other. These methods rely on the idea that words with similar meanings tend to occur in similar contexts (Harris, 1954). This proposition is called *distributional hypothesis* and provides a practical framework to understand and compute semantic relationship between words. Word embeddings have been used in a wide variety of applications such as sentiment analysis (Socher, Huval, Manning, & Ng, 2012), psychiatry (Bedi et al., 2015), psychology (Elias Costa, Bonomo, & Sigman, 2009; Sagi, Diermeier, & Kaufmann, 2013), philology (Diuk, Slezak, Raskovsky, Sigman, & Cecchi, 2012), literature (Altszyler & Brusco, 2015), cognitive science (Landauer, 2007), finance (Galvez & Gravano, 2017) and social science (Carrillo, Cecchi, Sigman, & Analysis, 2015; Kulkarni, Al-Rfou, Perozzi, & Skiena, 2015).

Latent Semantic Analysis (LSA) (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990; Hu, Cai, Wiemer-Hastings, Graesser, & McNamara, 2007; Landauer & Dumais, 1997), is one of the most used methods for word meaning representation. LSA takes as input a training corpus, i.e. a collection of documents. A word by document co-occurrence matrix is constructed. Typically, tf-idf transformation is applied to reduce the weight of uninformative high-frequency words in the words-documents matrix (Dumais, 1991). The output of the tf-idf transformation is a matrix W where the element $w_{i,j}$ is the weight of word i in the document j,

$$w_{i,j} = tf_{i,j} \cdot \log_2\left(\frac{D}{df_i}\right), \tag{1}$$

where $tf_{i,j}$ is the frequency of the word i in the document j,D is the number of documents in the training corpus and df_i is the number of documents in which word i appears. Then, each document weight is normalized to unit length and a dimensionality reduction is implemented by a *truncated Singular Value Decomposition*, where only the k largest singular values are selected. This method, provides a low-dimensional vectorial representation of every word present in the trained corpus. The success of LSA in capturing the latent meaning of words comes from this low-dimensional mapping (Turney & Pantel, 2010).

More recently, neural-network language embeddings have received increasing attention (Collobert & Weston, 2008; Mikolov, Chen, Corrado, & Dean, 2013), leaving aside classical word representation methods such as LSA. In particular, Word2vec models (Mikolov, Chen, et al., 2013; Mikolov et al., 2013) have become especially popular in embeddings generation.

Word2vec consists of two neural network language models, Continuous Bag of Words (CBOW) and Skip-gram. In both models, a window of predefined length is moved along the corpus, and in each step the network is trained with the words inside the window. Whereas the CBOW model is trained to predict the word in the center of the window based on the context words (the surrounding words), the Skip-gram model is trained to predict the context words based on the central word. In the present study, we use a Skip-gram model, which shows better performance in Mikolov, Corrado, et al. (2013) and in Asr, Willits, and Jones (2016) semantic tasks.

For each training example, the Skip-gram model tends to maximize the log probability of observed context words w_k given the center word w_j,

$$\log p(w_k|w_j) = \log\frac{\exp(\mathbf{c_k}\cdot\mathbf{v_j})}{\sum_i \exp(\mathbf{c_i}\cdot\mathbf{v_j})}, \tag{2}$$

where $\mathbf{c_k}$ and $\mathbf{v_j}$ are the vectorial representations of the context and central words, respectively, and the index i is spanning along all words in the vocabulary. Once the neural network has been trained, the average between both learned vectorial representations is taken as the final word representation. In order to increase the efficiency of the method, Mikolov, Chen, et al. (2013) propose a different version of the technique, the *negative sampling* method. In this variant, for each training example (w_k,w_j), the model is feed with a predefined number of words sampled from the vocabulary, as examples of words that did not appear in the context of w_j (for a more detailed explanation see (Jurafsky and Martin, 2014; Mikolov, Chen, et al., 2013)).

An intrinsic difference between LSA and Word2vec is that while LSA is a counter-based model, Word2vec is a prediction-based model. Although prediction-based models have strongly increased in popularity, it is not clear whether they outperform classical counter-based models (Baroni, Dinu, & Kruszewski, 2014; Levy & Goldberg, 2014; Levy, Goldberg, & Dagan, 2015).

In particular, Word2vec methods have a distinct advantage in handling large datasets, since they do not consume as much memory as some classic methods like LSA and, as part of the Big Data revolution, Word2vec has been trained with large datasets of about billions of tokens. However, only a few studies analyze semantic representations of small corpora, such as the typical dream collection. In a recent study, Asr et al. (2016) show that a co-occurrence model, like LSA, outperforms Skip-gram model in a semantic classification task over a medium size corpus (8 million words). In the same line, Sahlgren and Lenci (2016) compare the performance