Cognition 164 (2017) 46–60

Contents lists available at ScienceDirect

Cognition

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

Original Articles

The semantic representation of prejudice and stereotypes

Sudeep Bhatia

Department of Psychology, University of Pennsylvania, Philadelphia, PA, United States

ARTICLE INFO

Article history:
Received 22 March 2016
Revised 23 March 2017
Accepted 24 March 2017

Keywords:
Semantic representation
Latent semantic analysis
Prejudice
Stereotyping
Implicit association test

ABSTRACT

We use a theory of semantic representation to study prejudice and stereotyping. Particularly, we consider large datasets of newspaper articles published in the United States, and apply latent semantic analysis (LSA), a prominent model of human semantic memory, to these datasets to learn representations for common male and female, White, African American, and Latino names. LSA performs a singular value decomposition on word distribution statistics in order to recover word vector representations, and we find that our recovered representations display the types of biases observed in human participants using tasks such as the implicit association test. Importantly, these biases are strongest for vector representations with moderate dimensionality, and weaken or disappear for representations with very high or very low dimensionality. Moderate dimensional LSA models are also the best at learning race, ethnicity, and gender-based categories, suggesting that social category knowledge, acquired through dimensionality reduction on word distribution statistics, can facilitate prejudiced and stereotyped associations.

1. Introduction

Distributional models of semantic memory provide a powerful approach to understanding semantic representations (Griffiths, Steyvers, & Tenenbaum, 2007; Jones & Mewhort, 2007; Kwantes, 2005; Landauer & Dumais, 1997; Lund & Burgess, 1996; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Pennington, Socher, & Manning, 2014). One of the main insights underlying these models is that the representations of words reflect the structure of word co-occurrence in natural language (Firth, 1957; Harris, 1954). Studying this structure, by applying these models to large-scale natural language corpora, can shed light on the representations that people have of common words, the relationships and associations between the concepts that these words represent, and the ways in which these relationships affect cognition and behavior.

Distributional models often characterize the words in their vocabulary as multi-dimensional vectors, with the proximity between the vectors of two words corresponding to the relatedness or association of the words. The dimensionality of these vectors is often smaller than that necessary to represent the data on which the model is trained, so that learning the vector representations involves performing some type of dimensionality reduction on word distribution statistics (Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998). Appropriate levels of vector dimensionality allow distributional models to accurately predict response probabilities and response times in a wide range of settings, including semantic priming tasks, free association tasks, recall tasks, word similarity tasks, and categorization tasks (see Bullinaria & Levy, 2007 or Jones, Willits, & Dennis, 2015 for a review).

The use of distributional models is typically limited to non-social psycholinguistic settings. We wish to use these models to better understand prejudice and stereotyping. In this paper, we recover race-based, ethnicity-based, and gender-based vector representations from the types of natural language environments individuals interact with on a day-to-day basis, and examine whether our recovered representations possess the prejudiced and stereotyped associations documented in social psychological research. Importantly, we test the effects of mechanisms like dimensionality reduction on the strength of these prejudices and stereotypes. These mechanisms are necessary for the efficient learning of word meaning and association, and play a key role in the learning of categories. Examining whether these otherwise desirable cognitive mechanisms also generate undesirable social biases, can shed light on the cognitive underpinnings of these biases, and the ways in which these biases depend on social category knowledge and category-based generalization.

1.1. Prejudiced and stereotyped associations

Prejudice and stereotyping are often studied in terms of the associations that automatically influence judgment and behavior when relevant social categories are activated (Allport, 1954; Devine, 1989; Fazio, Jackson, Dunton, & Williams, 1995; Gaertner

E-mail address: bhatiasu@sas.upenn.edu

http://dx.doi.org/10.1016/j.cognition.2017.03.016
0010-0277/

& McLaughlin, 1983; Greenwald & Banaji, 1995; Strack & Deutsch, 2004). These associations are often considered to be implicit, that is, outside of the awareness of the individual in consideration. For this reason, they are studied using experimental tasks with measures that do not rely on the individual's ability to consciously assess (and suppress) these associations. Perhaps the most common such task in use today is the implicit association test (IAT) (Cunningham, Preacher, & Banaji, 2001; Greenwald, McGhee, & Schwartz, 1998), which provides a latency-based measure of associations for social categories. With the use of the IAT and related measures (Fazio & Olson, 2003), researchers have found stronger associations between stereotypically African American names and negatively valenced words and stronger associations between stereotypically White names and positively valenced words (Greenwald et al., 1998; also Dovidio, Evans, & Tyler, 1986; Fazio et al., 1995; Gaertner & McLaughlin, 1983), illustrating associative prejudices favoring Whites over African Americans. Similar methods have also been applied to study stereotypes, which do not involve diverging associations with differently valenced words, but rather diverging associations with words in different semantic categories. For example, researchers have used the IAT to demonstrate a stronger association between female names and weakness-related words and a stronger association between male names and power-related words (Rudman, Greenwald, & McGhee, 2001; also Nosek, Banaji, & Greenwald, 2002).

Biased associations have been shown to play a role in influencing peoples' behaviors (Dovidio, Kawakami, & Gaertner, 2002; Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Hamilton & Gifford, 1976; Judd & Park, 1988; McConnell & Leibold, 2001; Olson & Fazio, 2001; but also see e.g. Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013) and are considered to be one of the most important psychological determinants of prejudice and stereotyping. Given this importance, it becomes desirable to characterize what these associations are and the ways in which these associations are represented. One way to do this involves studying the distribution of names, words, and concepts in real-world natural language environments. People exposed to everyday language that presents African American names in negative contexts and White names in positive contexts, or female names in positions of weakness and male names in positions of power, will develop the prejudices and stereotypes documented in the above work. Equivalently, these prejudices and stereotypes will be reflected in the use of this language, causing African American names to be more likely to appear in negative contexts and less likely to appear in positive contexts, relative to White names, and female names to be more likely to appear in positions of weakness and less likely to appear in positions of power, relative to male names.

Studying the types of race-based or gender-based associations present in everyday language can not only shed light on the actual associations possessed by individuals, but also the ways in which these associations reflect social representations. This can then help us directly compare what we know about the cognitive basis of prejudice and stereotyping with what we know about the representation of non-social concepts. Does the representation of prejudice and stereotypes rely on same mechanisms involved in the representation of word relationships, categories, meanings, and associations in other settings? These mechanisms often facilitate efficient linguistic comprehension and word use, so could it be that prejudice and stereotyping are the harmful byproducts of an otherwise desirable system for making semantic inferences and generalizations?

1.2. Latent semantic analysis

These questions can be answered by applying theories of distributional semantics to common natural language datasets. The representations built using this method can then be tested for the types of associative biases observed in human participants, using, for example, stimuli from existing implicit association tests. The distributional model we consider in this paper is latent semantic analysis (LSA). LSA has been shown to be useful for a number of different applications in semantic memory research and computational linguistics, and is perhaps the most influential such model in this area (Landauer & Dumais, 1997; Landauer et al., 1998). Its core assumption is that decision makers represent words and concepts using a multidimensional word-vector space, built from word-distribution data. This vector space may have a high number of dimensions, but importantly, these dimensions are much less than those required for representing all of the information in the data. LSA achieves this dimensionality reduction using singular value decomposition.

Consider a setting with N different words occurring in K different contexts. These contexts could be different articles in newspapers, as in the dataset we consider below, chapters in books, conversations on the internet, or even non-textual experiences. The distribution of these words across the different contexts can be represented in an $N \times K$ matrix S. S captures word-context co-occurrence, so that the cell in row n and column k corresponds to the number of times word n occurs in context k.

LSA attempts to recover vector representations of the N words by performing a singular value decomposition on the matrix S, which describes S using some $M \ll K$ latent dimensions. The matrix recovered through this singular value decomposition can be written as $S^* = U \cdot V \cdot W$ where V is an $M \times M$ matrix with the M largest singular values from the decomposition, U is the corresponding $N \times M$ matrix of words, and W is the corresponding $M \times K$ matrix of contexts. U is of particular interest to us as it contains a representation of each of the N words as vectors on the M latent dimensions. The proximity between these vectors can be used to provide a quantitative account of word relationship and association. The metric typically used to compute vector proximity, and thus word association, is cosine similarity, so that the proximity between any two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by $\text{sim}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x} \cdot \boldsymbol{y} / (||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||)$. This metric varies between -1 and $+1$ (with 0 capturing orthogonal vectors and $+1$ capturing vectors with identical directions) (see Landauer et al., 1998 for details).

In their classic article, Landauer and Dumais (1997) showed that the above technique could be used to model judgments of word similarity and their dependence on the rate of vocabulary acquisition, specify the comprehension and comprehensibility of pieces of text, predict word priming effects, learn the representation of numerals, and display desirable properties in a number of other settings. Related work has shown that similar approaches are also able to predict human behavior in free association tasks, recall tasks, semantic categorization tasks, and in a wide variety of other psycholinguistic experiments (see Bullinaria & Levy, 2007; Jones et al., 2015; Turney & Pantel, 2010).

1.3. Dimensionality reduction

Importantly these results rely on the appropriate choice of M, which is the total number of latent dimensions recovered by singular value decomposition. Dimensionality reduction facilitates induction and generalization, so that if M is too large or if $M = K$ (which is the special case with no dimensionality reduction) the model is unable to generalize what it learns about words in a certain context to other word and other contexts. Thus, although an LSA model may note that the words *car* and *gear* occur together and are related, and the words *car* and *brake* occur together and are related, unless *gear* and *brake* occur together, an LSA model with a large value of M would not be able to infer that *gear* and *brake* are related. A very small value of M, or too much dimension-