Journal of Memory and Language 92 (2017) 57–78

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation

Paweł Mandera *, Emmanuel Keuleers, Marc Brysbaert

Department of Experimental Psychology, Ghent University, Belgium

ARTICLE INFO

Article history:
Received 10 June 2015
revision received 4 April 2016

Keywords:
Semantic model
Distributional semantics
Semantic priming
Psycholinguistic resource

ABSTRACT

Recent developments in distributional semantics (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) include a new class of prediction-based models that are trained on a text corpus and that measure semantic similarity between words. We discuss the relevance of these models for psycholinguistic theories and compare them to more traditional distributional semantic models. We compare the models' performances on a large dataset of semantic priming (Hutchison et al., 2013) and on a number of other tasks involving semantic processing and conclude that the prediction-based models usually offer a better fit to behavioral data. Theoretically, we argue that these models bridge the gap between traditional approaches to distributional semantics and psychologically plausible learning principles. As an aid to researchers, we release semantic vectors for English and Dutch for a range of models together with a convenient interface that can be used to extract a great number of semantic similarity measures.

Introduction

Distributional semantics is based on the idea that words with similar meanings are used in similar contexts (Harris, 1954). In this line of thinking, semantic relatedness can be measured by looking at the similarity between word co-occurrence patterns in text corpora. In psychology, this idea inspired a fruitful line of research starting with Lund and Burgess (1996) and Landauer and Dumais (1997). The goal of the present paper is to incorporate a new family of models recently introduced in computational linguistics and natural language processing research by Mikolov, Chen, Corrado, and Dean (2013) and Mikolov, Sutskever, Chen, Corrado, and Dean (2013) into psycholinguistics. In order to do so, we will discuss the theoretical foundation of these models and evaluate their performance on predicting behavioral data on psychologically relevant tasks.

Count and predict models

Although there are different approaches to distributional semantics, what they have in common is that they start from a text corpus and that they often represent words as numerical vectors in a multidimensional space. The relatedness between a pair of words is quantified by measuring the similarity between the vectors representing these words.

The original computational models of semantic information (arising from the psychological literature) were based on the idea that the number of co-occurrences of words in particular contexts formed the basis of the multi-

* Corresponding author at: Ghent University, Department of Experimental Psychology, Henri Dunantlaan 2, Room 150.025, 9000 Ghent, Belgium.
E-mail address: pawel.mandera@ugent.be (P. Mandera).

http://dx.doi.org/10.1016/j.jml.2016.04.001
0749-596X/

dimensional space and that the vectors were obtained by applying a set of transformations to the count matrix. For instance, Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) starts by counting how many times a word is observed within a document or a paragraph. The Hyperspace Analogue to Language (HAL; Lund & Burgess, 1996) counted how many times words co-occurred in a relatively narrow sliding window, usually consisting of up to ten surrounding words. Because of the common counting step, following Baroni, Dinu, and Kruszewski (2014) we will refer to this family of models as ***count models***.

In count models, the result of this first step is a word by context matrix. What usually follows is a series of transformations applied to the matrix. The transformations involve some kind of a weighting scheme, based on frequency-inverse document frequency, positive pointwise mutual information (PPMI), log-entropy, and/or a dimensionality reduction step (most commonly singular value decomposition; SVD). Sometimes the transformation is the defining component of the method, as is the case for LSA, which is based on SVD. In other cases, however, the transformations have been applied rather arbitrarily to the counts matrix based on empirical studies investigating which transformations optimized the performance on a set of tasks. For example, in its original formulation, the HAL model did not involve complex weighting schemes or dimensionality reduction steps, but later it was found that they improved the performance of the model (e.g., Bullinaria & Levy, 2007, 2012). Transformations are now often applied when training the models (e.g., Mandera, Keuleers, & Brysbaert, 2015; Recchia & Louwerse, 2015).

If we consider Marr's (1982) distinction between computational, algorithmic, and implementational levels of explanation, the count models are *only defined* at the *computational* level (Landauer & Dumais, 1997, p. 216): They consist of functions that map from a text corpus to a count matrix and from the count matrix to its transformed versions. Regarding the algorithmic level, Landauer and Dumais (1997) did not attribute any realism to the mechanisms performing the mapping. They only proposed that the counting step and its associated weighting scheme could be seen as a rough approximation of conditioning or associative processes and that the dimensionality reduction step could be considered an approximation of a data reduction process performed by the brain. In other words, it cannot be assumed that the brain stores a perfect representation of word-context pairs or runs complex matrix decomposition algorithms in the same way as digital computers do.[1] In the case of HAL, even less was said about the psychological plausibility of the selected algorithms. Another problem is that count models require all the information to be present before the transformations are applied, whereas, in reality, learning in cognitive systems is incremental, not conditional on the simultaneous availability of all information.

In other words, although the count models, like all computational models, were very specific about which properties were extracted from the corpus to build the count matrix, and which mathematical functions were applied to the counts matrix in the transformation step, they made it much less clear how these computations could be performed by the human cognitive system.[2] This is surprising, given that the models originated in the psychological literature.

Unexpectedly, a recent family of models, which originated in computer science and natural language processing, may be more psychologically plausible than the count models. Mikolov, Chen, et al. (2013) argued that a relatively simple model based on a neural network (see Fig. 1) can be surprisingly efficient at creating semantic spaces.

This family of models is built on the concept of prediction. Instead of explicitly representing the words and their context in a matrix, the model is based on a relatively narrow window (similar in size to the one often used in the HAL model) sliding through the corpus. By changing the weights of the network, the model learns to predict the current word given the context words (Continuous Bag of Words model; CBOW) or the context words given the current word (skip-gram model). Because of the predictive component in this family of models, again following Baroni et al. (2014), we will refer to these models as ***predict models***. As indicated above, there are two main types: the CBOW model and the skip-gram model. Even though the predict models originated outside the context of psychological research and were not concerned with psychological plausibility, the simple underlying principle – implicitly learning how to predict one event (a word in a text corpus) from associated events – is arguably much better grounded psychologically than constructing a count matrix and applying arbitrary transformations to it. The implicit learning principle is congruent with other biologically inspired models of associative learning (Rescorla & Wagner, 1972), given that they both learn on the basis of the deviation between the observed event and the predicted event (see Baayen, Milin, Filipovic Durdevic, Hendrix, & Marelli, 2011). An additional advantage of the model is that it is trained using a stochastic gradient descent, which in this case means that it can be trained incrementally with only one target–context pairing available for each update of the weights, and does not require all co-occurrence information to be present simultaneously as is the case with the count models.

To illustrate in what sense we consider the predict models to be psychologically plausible, we would like to compare them to the Rescorla–Wagner model – a classical learning model (for a review see Miller, Barnet, & Grahame, 1995), which has also been successfully applied to psycholinguistics (Baayen et al., 2011). This model learns to associate cues with outcomes by being sequentially presented with training cases. For each training case, if there is a discrepancy between the outcomes predicted based

[1] It is known that dimension reduction can be performed by biological (e.g. Olshausen & Field, 1996) and artificial (Hinton & Salakhutdinov, 2006) neural networks. This fact is rarely mentioned when authors discuss various approaches to distributional semantics in the psycholinguistic literature.

[2] Although Landauer and Dumais (1997) discuss how the LSA algorithm could hypothetically be implemented in a neural network, this aspect is not reflected in their implementation of the model.