Title
Evaluating the Contribution of Intra-Linguistic and Extra-Linguistic Data to the Structure of Human Semantic Representations

Permalink
https://escholarship.org/uc/item/9ss6k1pt

Journal

ISSN
1069-7977

Authors
Andrews, Mark
Vigliocco, Gabriella
Vinson, David

Publication Date
2007

Peer reviewed
Evaluating the Contribution of Intra-Linguistic and Extra-Linguistic Data to the Structure of Human Semantic Representations

Mark Andrews (m.andrews@ucl.ac.uk)
Gabriella Vigliocco (g.vigliocco@ucl.ac.uk)
David Vinson (d.vinson@ucl.ac.uk)
Department of Psychology, University College London,
26 Bedford Way
London, WC1H 0AP
United Kingdom

Abstract

We describe Bayesian models that learn semantic representations from either extra-linguistic data or intra-linguistic data, or from both in combination. We evaluate the validity of these models using three human-based measures of semantic similarity. The results provide strong evidence for the hypothesis that human semantic representations are the product of the statistical combination of extra- and intra-linguistic sources of data.

Introduction

For the purposes of this paper, we use the term semantic representation to refer to a language user’s mental or cognitive representation of the meaning of words. We formally define this as the knowledge that allows the language user to infer, amongst other things, which are the semantic or ontological categories to which a word belongs, what (if anything) are the referents of a word. Our general aim is to consider how both extra-linguistic and intra-linguistic data can be used to acquire this knowledge. Extra-linguistic, or attributional data, is data that is derived from our perception and interaction with the physical world, and in particular, from the perceived physical attributes or properties associated with the referents of words. In contrast, intra-linguistic, or distributional data, is data that is derived from the statistical characteristics within a language itself, or how a given word is distributed across different spoken or written texts.

In previous literature, it has been repeatedly demonstrated that semantic representations can be learned from either attributional data alone, e.g., McRae, Sa, and Seidenberg (1997); Vigliocco, Vinson, Lewis, and Seidenberg (1997); Vigliocco, Vinson, Lewis, and Seidenberg (2002). However, in previous work of our own (Andrews, Griffiths and Steyvers (2002), we considered the combined effects of both extra-linguistic and distributional data, or from both in combination. We then evaluate the validity of these models using three human-based measures of semantic similarity: word-association norms, semantic-priming results from a lexical decision task, and interference patterns from a picture-word interference task.

Model Description

We provide Bayesian models that learn semantic representations from examples of attributional data, or from distributional data, or from both combined. The probabilistic models we employ for each of the various data types are described graphically in Figure 1. The attributional model (leftmost) describes any given word \( w_f \) as a probability distribution over a set of binary attributes, such that \( \{y_{m[f]} : 1 \leq m \leq M[f]\} \) is a set of bit vectors, each being an instance of the referent of the word \( w_f \). These probability distributions are compositions of a basic repertoire of latent distributions \( \psi = \{\psi_1 \ldots \psi_{K_{\text{Lat}}}\} \) that intuitively correspond to clusters of interrelated attributes each describing basic characteristics of the attributional data. The distributional model (second left) describes texts as multinomial distribution over words, such that \( \{w_{n[t]} : 1 \leq n \leq N[t]\} \) is a sample of words from text \( t \). These distributions are compositions of latent distributions \( \phi = \{\phi_1 \ldots \phi_{K_{\text{Disc}}}\} \) that intuitively correspond to discourse-topics in a corpus of text. The combined model (second right) describes texts as probability distributions over words, and words as distributions over attributes. These distributions

---

1For example, the word apple refers to objects in the world whose perceived attributes or properties include being red or green, round, shiny, smooth, crunchy, juicy, sweet, tasty, etc.

2We use the term text here in a very general sense to refer to any coherent and self-contained piece of written or spoken language. This could include, for example, a newspaper article, a spoken conversation, a letter or email message, an essay, a speech, etc.
In total there were 7818 unique word types.

Semantic Representations

The latent distributions in each model intuitively correspond to that model’s semantic knowledge. We can provide examples of this knowledge by drawing samples from the mean of the posterior distribution over the latent distributions. Examples for the cases of the attributional, distributional and combined models are shown in Table 1(a). From these examples, it can be seen that the latent-distributions are clusters of inter-related attributes (in the case of the attributional model), or words (in the distributional model), or both (in the combined model). Importantly, in the case of the combined model, attribute clusters align with discourse-topics that are consistent with the same general meaning.

Within each model, each word can be expressed as a distribution over that model’s latent distributions. From this we can measure the correspondence between any pair of words in each model. In general, in a model whose unobserved indicator variable is denoted by \( x \) the correspondence between words \( w_i \) and \( w_j \) is given by

\[
\text{Corr}(w_i, w_j) = \sum_{x} P(w_i | x) P(w_j | x). 
\]

In Table 1(b), using this formula, and averaging over samples from the posterior over the parameters, we provide examples of the near-neighbors of a set of example words according to each of our four models.

Model Evaluation

We evaluate each model by comparing its set of inter-word similarities with human-based measures of semantic similarity. There is, of course, no flawless means by which to measure human semantic representations or the inter-word similarities implied by them. In light of this, we have used a collection of methods that will hopefully lead to converging evidence. These are the Nelson word-association norms, and semantic-priming re-

---

3Each text was approximately 200-250 words in length. In total there were 7818 unique word types.

---

4We do not display examples from the independent model as these will, by design, be identical to the independent product of the attributional and distributional models' latent distributions.

5http://w3.usf.edu/FreeAssociation/
Table 1: Semantic Knowledge and Inter-word Similarities in the Models

(a) Examples of latent distributions learned by the attributional model (upper left), distributional model (lower left) and combined model (right). The latent distributions in the combined model are coupled distributions over both attributes and words.

(b) Examples of the near neighbors of a set of five words (boldface) according to the attributional, distributional, combined and independent models. The five words were chosen so as to highlight the differences between the four models.
Table 2: Interpretation of $\lambda$ in the Bayes Factor Test.

The analysis we will pursue is often referred to as the Bayes factor test. Given any test data-set $D_{\text{test}}$ and any two alternative models $M_0$ and $M_1$ (parameterized by $\theta_0$ and $\theta_1$, respectively) the Bayes factor for $M_1$ relative to $M_0$ is given by

$$
\lambda = \frac{P(D_{\text{test}}|M_1)}{P(D_{\text{test}}|M_0)} = \frac{\int d\theta_1 P(D_{\text{test}}|\theta_1)P(\theta_1|M_1)}{\int d\theta_0 P(D_{\text{test}}|\theta_0)P(\theta_0|M_0)}.
$$

The term $\lambda$ is a measure of evidence for the superiority of $M_1$ over $M_0$. Jeffreys (1961) provides a scale of interpretation for $\lambda$ as shown in Table 2. Clearly, this test is easily applied to our model comparisons, whereby we integrate over the posterior probabilities of the parameters for each model, evaluating the probability of data set for each parameter value. In our case, however, we must replace the integral with a sum over samples from the posteriors.

**Word Association Norms** The Nelson word association norm data-set is a collection of the close word-associates of 5019 English words. These have been collected from human participants under controlled circumstances, and each word associate is assigned a probability indicating the relative frequency of its being paired with the target word. Of the 5019 words, a subset of 2824 instances, and each word associate is assigned a probability given prime-target word-pairs, all of which occur in our data-sets, were collected. The speed of response to the target word, given the presence of the prime, is compared to the speed of response of the target word in the presence of an obviously unrelated baseline word (matched to the prime word on salient characteristics such as length, frequency, etc.). This allows each prime-target pair to be represented in terms of the relative speed up of response to the target in the presence of the prime.

In order to assess how well each model predicts this data, we used a separate Bayesian linear regression model for each model. In each case, we regressed relative speed-up (msec) for target-word $v_i$ given prime-word $v_j$ on the log of $P(v_j|v_i)$ derived from each model. The quantity $P(v_j|v_i)$ was obtained by averaging over the posterior of the parameters in each model. The outcome of the Bayesian regression is a posterior distribution over the parameters of the linear-Gaussian regression model. From this, we can calculate the marginal likelihood of the priming-data for the case of each model, and compare these in a Bayes factor test as in Equation 1. We have plotted the log of these marginal likelihoods in the upper right sub-figure of Figure 2. Note that, as before, the differences between any pair of log probabilities will be equal to the log of the ratio of these probabilities. As can be seen in this figure, there is very strong evidence (using the Jeffreys' definition) in favor of the superiority of the combined model's predictiveness of the word association norms.

For the purposes of comparison, we performed an analysis of these data using non-parametric statistics and sampling based null hypothesis tests and the relative ordering of the models' performances was identical.

<table>
<thead>
<tr>
<th>$\log \lambda$</th>
<th>Evidence for $M_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\log \lambda &lt; 0$</td>
<td>Negative</td>
</tr>
<tr>
<td>$0 \leq \log \lambda &lt; 1$</td>
<td>Weak</td>
</tr>
<tr>
<td>$1 \leq \log \lambda &lt; 2.5$</td>
<td>Positive</td>
</tr>
<tr>
<td>$2.5 \leq \log \lambda &lt; 5$</td>
<td>Strong</td>
</tr>
<tr>
<td>$\log \lambda \geq 5$</td>
<td>Very strong</td>
</tr>
</tbody>
</table>

For the purposes of comparison, we performed an analysis of these data using non-parametric statistics and sampling based null hypothesis tests and the relative ordering of the models’ performances was identical.
For the purpose of comparison, it is also useful to consider the results from a standard, or non-Bayesian, linear regression test. Commonly used measures from this type of analysis include a measure of the strength of the linear relationship between the variables $R$, the amount of variance in the dependent variable accounted for by the independent variable $R^2$, and the $p$-value significance of these statistics $p$. These are as follows:

<table>
<thead>
<tr>
<th>Model</th>
<th>$R$</th>
<th>$R^2$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attributional</td>
<td>.31</td>
<td>.09</td>
<td>.006</td>
</tr>
<tr>
<td>Distributional</td>
<td>.29</td>
<td>.086</td>
<td>.008</td>
</tr>
<tr>
<td>Combined</td>
<td>.39</td>
<td>.16</td>
<td>.0002</td>
</tr>
<tr>
<td>Independent</td>
<td>.22</td>
<td>.05</td>
<td>.04</td>
</tr>
<tr>
<td>Unigram</td>
<td>.078</td>
<td>.006</td>
<td>.49</td>
</tr>
</tbody>
</table>

**Picture Word Interference** In a picture-word interference task, naming latencies of drawings of objects (or actions/events) are recorded. When these pictures are presented simultaneously with a word, and if that word is semantically related to the picture, naming latencies increase. This increase resembles the Stoop phenomenon whereby the semantically related word interferes with the activation of the picture’s name. In Vigliocco et al. (2004), picture-word interference data was collected for a set of word pairs (all of which occur in all our models). If the picture depicts word $v_i$ and the distractor word is $v_j$, the slow-up for naming the picture as $v_i$ (relative to a baseline distractor) can be used as a measure of the semantic similarity between $v_i$ and $v_j$.

As in the case of the priming data, we used separate Bayesian linear regression models, regressing naming latency against the log of $P(v_i|v_j)$ in each model (averaging over parameters). From this, we can calculate the marginal likelihood of the picture-word interference data according to each model. The marginal likelihoods can be compared in a Bayes factor test, as before. We have plotted the log of these marginal likelihoods in the lower right sub-figure of Figure 2. The relative pattern of results is almost identical to that seen in the priming data case. The combined model shows the strongest predictive power with the ordering from strongest to weakest model is combined model, distributional model, independent model, attributional and unigram models. These results are strongly supported by the Bayes factor test.

As with the case of the priming data, for the purposes of comparison, we can mention standard measures from non-Bayesian regression analysis, i.e. $R$, $R^2$ and $p$:

<table>
<thead>
<tr>
<th>Model</th>
<th>$R$</th>
<th>$R^2$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attributional</td>
<td>−.24</td>
<td>.06</td>
<td>.09</td>
</tr>
<tr>
<td>Distributional</td>
<td>−.35</td>
<td>.12</td>
<td>.01</td>
</tr>
<tr>
<td>Combined</td>
<td>−.38</td>
<td>.14</td>
<td>.009</td>
</tr>
<tr>
<td>Independent</td>
<td>−.26</td>
<td>.06</td>
<td>.08</td>
</tr>
<tr>
<td>Unigram</td>
<td>−.19</td>
<td>.03</td>
<td>.19</td>
</tr>
</tbody>
</table>
Discussion

The general aim in this paper has been to consider how semantic representations are acquired. To answer this we have identified two major types of data from which semantic information can be attained. We have referred to these as attributional and distributional data types. These represent data types that are, respectively, extra-linguistic and intra-linguistic in their origin. Of particular concern to us has been the question of how these two distinct data types can be combined to learn coherent semantic representations. We have provided a model of the semantic representations that are learned from attributional and distributional data taken in combination, and compared this to the representations learned from either source taken independently. Our specific aim has then been to evaluate these models against human-based measures of semantic representations.

Although the relative performance of each model to predict the human data is not identical across the three different data-sets there are obvious and compelling general trends. For example, and unsurprisingly, all four of the attributional, distributional, combined and independent models outperform the null model on all data-sets. While superior performance against a null-model is not surprising, it does serve as a worthwhile sanity check, effectively corroborating the impression given by Table 1 that each of these models is providing (at the very least) a modest description of the meaning of words.

If the unigram model represents a lower-bound on the models’ predictive performances, then it appears as if the combined model represents an upper-bound. The combined model outperforms all other models consistently across all three sets of human-based measures. This corroborates the impression given by Table 1(b) that the combined model provides a more comprehensive and valid account of the meanings of words than do either the attributional, distributional or independent models. As such, we can take this as direct evidence in favor of our primary hypothesis that human semantic representations are the product of the statistical combination, and not simply the sum or average, of attributional and distributional data-types.

Conclusion

The results imply a certain picture of how word-meanings are learned. This can be described by reference to following scenario: A child learning his or her native language will regularly experience words referring to, for example, everyday objects in the context of one or more of their referents. On the other hand, the words that the child is learning are not necessarily heard in isolation, but rather will regularly occur in the context of meaningful sentences. From this, the data from which the child can learn word-meanings occur in two forms simultaneously: There is the set of attributes associated with a given word, and the set of textual contexts in which that word occurs. While it has been repeatedly shown in previous literature that either one of these sources can provide information from which word-meanings can be learned, learning from both data-types in combination would allow the correspondences between the two data-types to be apparent, and to be exploited. For example, if the child learned that the word cat refers to creatures with claws and whiskers and tails, etc. and that it also co-occurs with terms like dog, pet, owner, etc., it may also infer that creatures with claws and whiskers and tails, etc., are conceptually related to the words dog, pet, owner, etc. From this, we can see that while using either extra-linguistic or intra-linguistic data can allow semantic representations to be learned by discovering the correlations within that specific data-type, using the combination of both allows the discovery of correlations both within and between these data-types.

References


