A Metric Scale for ‘Abstractness’ of the Word Meaning

Alexei V. Samsonovich
Krasnow Institute for Advanced Study, George Mason University, Fairfax, Virginia, USA
asamsono@gmu.edu

Abstract
Web personalization involves automated content analysis of text, and modern technologies of semantic analysis of text rely on a number of scales. Among them is the abstractness of meaning, which is not captured by more traditional measures of sentiment, such as valence, arousal and dominance. The present work introduces a physics-inspired approach to constructing the abstractness scale based on databases of hypernym-hyponym relations, e.g., WordNet 3.0. The idea is to define an energy as a function of word coordinates that are distributed in one dimension, and then to find a global minimum of this energy function by relocating words in this dimension. The result is a one-dimensional distribution that assigns “abstractness” values to words. While positions of individual words on this scale are subject to noise, the entire distribution globally defines the universal semantic dimension associated with the notion of hypernym-hyponym relations, called here “abstractness”.

Keywords: semantic mapping; sentiment analysis; quantification of meaning; recommender systems.

Introduction
Many applications in web personalization, automated recommender systems, sentiment analysis, semantic search, and related domains involve automated evaluation of word semantics on various scales, or dimensions, that characterize the word meaning. These tools are also known as semantic maps or semantic spaces (Gardenfors 2004). Among the best-known models are the semantic differential introduced by Osgood (Osgood, Suci, and Tannenbaum 1957), that gave rise to a number of related models of semantic vector spaces known by different names, e.g., EPA (evaluation, potency, activity) or PAD (pleasure, arousal, dominance), the Circumplex model (Russell, 1980), and related databases (e.g., ANEW: Bradley and Lang, 1999). Since Osgood, is generally understood that the large variety of semantic scales and dimensions used in these studies fall into the few broadly defined principal semantic categories, or components. This understanding is also consistent with the recent analysis of synonym-antonym relations performed using an energy minimization approach (Samsonovich and Ascoli 2007, 2010). The semantic components identified in this study can be approximately characterized as “positive vs. negative” (valence), “strong vs. weak” (dominance, arousal), “open vs. closed”, etc. The present study extends the energy minimization approach to one new dimension that does not appear to be reducible to those listed above, and can be interpreted as abstractness of the word meaning.

Computational tools were recently developed (e.g., Coh-Metrix: McNamara et al. 2006) for inferring abstractness of words from a given hypernym-hyponym hierarchy. However, a perfect hierarchy is not always available, and the problem in general remains open. For example, the graph of hypernym-hyponym relations available in WordNet 3.0 includes a high density of cycles and other mutually inconsistent links, making a straightforward ordering of words based on these relations impossible. The present study uses a physics-inspired energy minimization approach to solving this problem.

The constructed scale for “abstractness” can be further improved using alternative approaches. It has implications for automated semantic analysis of text (Krippendorf 2004, Pang and Lee 2008), that is an essential component in web personalization and automated recommender systems.

Methods
The idea of the optimization approach used here is to construct an energy function of word coordinates that are allocated in one dimension, and then to find a global minimum of this energy by relocating words, which is done using standard optimization techniques. The energy function is constructed using hypernym-hyponym relations among words, with its analytical form derived from first principles. The resultant one-dimensional distribution of words defines a semantic scale for “abstractness”.

Copyright © 2012, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
Derivation of the Energy Function

Suppose we want to measure correlates of the semantic dimension associated with the hypernym-hyponym relation (call it “abstractness” of the meaning of words) in a given set of \( n \) words that are numbered from 1 to \( n \). In principle, this can be done directly, if we know a variable \( A_{ij} \) that takes values \{-1, 0, 1\} and is defined for a certain subset of word pairs \((i, j)\) as follows:

\[
A_{ij} = \begin{cases} 
1, & \text{if } i \text{ is a hypernym of } j, \\
-1, & \text{if } i \text{ is a hyponym of } j, \\
0, & \text{if the relation is absent or unknown.}
\end{cases}
\]

Given this definition of \( A \), we can agree that whenever \( A_{ij} = +1 \), then the meaning of \( i \) is more “abstract” than the meaning of \( j \), and vice versa: whenever \( A_{ij} = -1 \), then the meaning of \( i \) is less “abstract” than the meaning of \( j \). Therefore, we can measure correlates of thus understood “abstractness” of the word meaning, if we can measure correlates of \( A \) defined on the given set of word pairs.

Practically, in most cases it would be more convenient for us to measure correlates of “abstractness” by sampling individual words rather than word pairs. To do this, we need a variable \( x_i \) assigned to each word \( i \) that is a good measure of “abstractness”. Ideally, for this purpose we want \( x \) to be a best possible measure of “abstractness”, given the information available in \( \{A_{ij}\} \). This task can be defined mathematically precisely. Because we are interested in measuring correlates of \( A \) using \( x \), we want to use values of \( x \) that maximally correlate with \( A \). In other words, the values \( \{x_i\} \) should maximize the following Pearson correlation coefficient \( R \):

\[
R = \text{corr}_G \left( A_{ij}, x_i - x_j \right).
\]

where the subscript \( G \) means that only selected pairs of words are used (see below), and the Pearson correlation is

\[
\text{corr}(x, y) = \frac{\sum (x_i y_i - \bar{x} \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}.
\]

In our case, all three sums in \( R \) are computed over those pairs of words for which \( A \) is nonzero. This set of ordered pairs of words is labeled \( G \) (note that if \( G \) contains a pair \((i, j)\), then it also contains the pair \((j, i)\)):

\[
G = \{(i, j) : A_{ij} \neq 0\}.
\]

Here we prefer to miss some information that may be partially conveyed via zero \( A \)’s, exactly because we do not know how many of zero \( A \)’s correspond to actually missing relations. The computation of \( R \) is straightforward:

\[
R = \text{corr}_G \left( A_{ij}, x_i - x_j \right)
= \frac{\langle A_{ij}(x_i - x_j) \rangle_G}{\langle (A_{ij} - \bar{A})^2 \rangle_G^{1/2} \langle (x_i - x_j)^2 \rangle_G^{1/2}}.
\]

Here the angular brackets \( \langle \ldots \rangle_G \) mean the average computed over all ordered word pairs in \( G \), and we took into account that the average of the differences \( x_i - x_j \) over all word pairs in \( G \) is zero. Also, the average \( \bar{A} = \langle A_{ij} \rangle_G \) is zero by the definition of \( A \), and the absolute values of all \( A_{ij} \)'s are 1. Therefore, the expression for \( R \) simplifies, and the problem can be formulated as follows:

\[
\bar{x} = \arg\max_{\bar{x} \in \mathbb{R}^n} \frac{\langle A_{ij}(x_i - x_j) \rangle_G}{\langle (x_i - x_j)^2 \rangle_G^{1/2}}. \tag{1}
\]

In other words, we want to find an \( \bar{x} \) that maximizes \( R \). The problem (1) is not very convenient for a numerical solution. The function is nonlinear, has singularities, and in addition is scale-invariant: if \( x \) is a solution of (1), then \( \alpha x \) is also a solution of (1) for any \( \alpha \neq 0 \). In order to define the problem (1) satisfactorily for a numerical solution, we need to add a constraint on \( x \) that eliminates the scale invariance. For example, we can require that the standard deviation of the differences (the square root in the denominator of (1)) is equal to 1. Thus, (1) is reduced to a constrained optimization problem:

\[
\begin{cases}
\bar{x} = \arg\max_{\bar{x} \in \mathbb{R}^n \setminus \{0\}} \langle A_{ij}(x_i - x_j) \rangle_G, \\
\langle (x_i - x_j)^2 \rangle_G = 1.
\end{cases} \tag{2}
\]

Now, using the method of Lagrange multipliers (e.g., Bertsekas, 1999), a necessary condition for a solution of (2) can be written as follows:

\[
\bar{x} = \arg\max_{\bar{x} \in \mathbb{R}^n} \left( \lambda_1 \sum_G (x_i - x_j)^2 - \sum_G A_{ij} (x_i - x_j) \right), \tag{3}
\]

where \( \lambda_1 \) is some constant (not necessarily positive). This problem (3) is in turn equivalent to the problem

\[
\bar{x} = \arg\max_{\bar{x} \in \mathbb{R}^n} \sum_G \left[ (x_i - x_j)^2 - 2\lambda A_{ij} (x_i - x_j) + \lambda^2 A_{ij}^2 \right], \tag{4}
\]

where \( \lambda \) is another constant related to \( \lambda_1 \); finally, (4) is equivalent to

\[
\bar{x} = \arg\min_{\bar{x} \in \mathbb{R}^n} \sum_G (x_i - x_j - \lambda A_{ij})^2. \tag{5}
\]
We observe that:

- the constant term in the bracket in (5) only affects the scale of \( x \) but not the correlation coefficient, and therefore this term can be replaced with 1;
- in the sum of (5), for each term with a positive \( A_{ij} \) there is a matching term with the negative \( A_{ij} \) that adds exactly the same contribution to the sum.

Therefore, we can eliminate one half of the terms from the sum, rewriting (5) as follows:

\[
\bar{x} = \arg\min_{\mathbb{R}^n} \sum_{i,j=1}^{n} W_{ij} (x_i - x_j - 1)^2,
\]

where the selection of terms in the sum is actually done by the new variable \( W \) defined as

\[
W_{ij} = \begin{cases} 
1, & \text{if } A_{ij} = 1, \\
0 & \text{otherwise.}
\end{cases}
\]

Because (6) is a necessary condition, and the function under the argmin (call it the energy function) has exactly one minimum (because the function is a quadratic form), this minimum yields the solution of the original problem: the value of \( \bar{x} \) defined by (6) has the maximal possible correlation with \( A \) (however, see the Remark below).

The solution of (6) is still translationally invariant. Therefore, one can add an arbitrarily scalar constant to \( x \). Moreover, if the graph of hyponym-hypernym links is not connected, then arbitrary, different scalar constants can be added to the values of \( x \) in disjoined graph components, which means that the variance of the distribution of \( x \)’s can grow without control during optimization. Practically, this means that a problem may emerge during the optimization process even in cases when the graph is fully, yet weakly connected. E.g., weakly connected, randomly diluted parts of the graph may end up on the periphery of the distribution, far away from the meaningful values of \( x \).

Due to these and other reasons, the “energy function” in (6) needs to be regularized. One way to do this without complicating the problem is by adding a simplest regularizing quadratic term, e.g., in the following form:

\[
\bar{x} = \arg\min_{\mathbb{R}^n} \left[ \sum_{i,j=1}^{n} W_{ij} (x_i - x_j - 1)^2 + \mu \sum_{i=1}^{n} x_i^2 \right],
\]

where \( \mu \) is the regularization parameter that should be adjusted in order to serve its purpose of stabilization of the optimization process and not distort the map significantly.

Remark: The above logic assumed that components of \( \{A_{ij}\} \) are statistically mutually independent variables, while in general they may not be independent. This implies that the formula (7) can be interpreted as an approximation, while a more detailed study would be necessary to produce a more accurate definition of the “abstractness” scale. This possibility needs to be explored elsewhere.

Materials and Procedures

The publicly available WordNet 3.0 database was used in this study. The hypernym-hyponym relations among \( n = 125,000 \) English words were extracted from the database as a connected graph defining the matrix \( W \), which was used to compute the energy function.

Optimization was done using standard Matlab functions and took several hours on a PC. The starting point for optimization was a random distribution of words on a line.

Results

The resultant distribution of words on thus computed scale of “abstractness” is represented in Figure 1. The two tails of the distribution of 125,000 words are given in Table 1.

![Figure 1. Histogram of the optimized distribution of words on the scale of “abstractness”.](image)

Examining the distribution in detail intuitively confirms the expectation that the order of words on the line roughly corresponds to the relative abstractness of their meaning. While positions of individual words on this scale are subject to noise, the entire distribution of words minimizes inconsistencies in the mapping of hypernym-hyponym links, as ensured by (7). Therefore, the final distribution (Figure 1) globally defines the “abstractness” dimension with the maximal accuracy afforded by the available data.

Preliminary analysis shows no strong correlations \((r<.2)\) between “abstractness” and other principal semantic dimensions, including valence and arousal, extracted from WordNet using a related optimization technique described previously (Samsonovich and Ascoli 2010). Therefore, thus defined “abstractness” appears to be an independent semantic characteristic. There are, however, indications that it may be weakly positively correlated with valence. This possibility needs further investigation elsewhere.
Table 1. The two tails of the distribution.

<table>
<thead>
<tr>
<th>“Abstractness” value</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.29237</td>
<td>entity</td>
</tr>
<tr>
<td>2.90877</td>
<td>physical entity</td>
</tr>
<tr>
<td>2.83684</td>
<td>psychological feature</td>
</tr>
<tr>
<td>2.69681</td>
<td>auditory communication</td>
</tr>
<tr>
<td>2.67756</td>
<td>unmake</td>
</tr>
<tr>
<td>2.64012</td>
<td>cognition</td>
</tr>
<tr>
<td>2.64012</td>
<td>knowledge</td>
</tr>
<tr>
<td>2.64012</td>
<td>noesis</td>
</tr>
<tr>
<td>2.61129</td>
<td>natural phenomenon</td>
</tr>
<tr>
<td>2.58016</td>
<td>ability</td>
</tr>
<tr>
<td>2.52437</td>
<td>social event</td>
</tr>
<tr>
<td>2.52384</td>
<td>craniate</td>
</tr>
<tr>
<td>2.52384</td>
<td>vertebrate</td>
</tr>
<tr>
<td>2.51277</td>
<td>higher cognitive process</td>
</tr>
<tr>
<td>2.50857</td>
<td>physiological property</td>
</tr>
<tr>
<td>2.47591</td>
<td>mammal</td>
</tr>
<tr>
<td>2.47591</td>
<td>mammalian</td>
</tr>
<tr>
<td>2.47445</td>
<td>denominate</td>
</tr>
<tr>
<td>2.47434</td>
<td>temporal property</td>
</tr>
<tr>
<td>2.45715</td>
<td>cerebration</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>-3.66013</td>
<td>green-blindness</td>
</tr>
<tr>
<td>-3.66013</td>
<td>protanopia</td>
</tr>
<tr>
<td>-3.66013</td>
<td>red-blindness</td>
</tr>
<tr>
<td>-3.76864</td>
<td>chain wrench</td>
</tr>
<tr>
<td>-3.79372</td>
<td>francis turbine</td>
</tr>
<tr>
<td>-4.18298</td>
<td>tricolour television tube</td>
</tr>
<tr>
<td>-4.18298</td>
<td>tricolour tube</td>
</tr>
<tr>
<td>-4.56639</td>
<td>edmontonia</td>
</tr>
<tr>
<td>-4.56639</td>
<td>coelophysis</td>
</tr>
<tr>
<td>-4.56639</td>
<td>deinocheirus</td>
</tr>
<tr>
<td>-4.56639</td>
<td>struthiomimus</td>
</tr>
<tr>
<td>-4.56639</td>
<td>deinonychus</td>
</tr>
<tr>
<td>-4.56639</td>
<td>dromeosaur</td>
</tr>
<tr>
<td>-4.56639</td>
<td>mononychus olecranus</td>
</tr>
<tr>
<td>-4.56639</td>
<td>oviraptorid</td>
</tr>
<tr>
<td>-4.56639</td>
<td>superslasher</td>
</tr>
<tr>
<td>-4.56639</td>
<td>utahraptor</td>
</tr>
<tr>
<td>-4.56639</td>
<td>velociraptor</td>
</tr>
</tbody>
</table>

Computational linguistics is an old field. The idea of defining and measuring abstractness of words based on word relations is by far not new. This is also true about the specific version of the idea that uses the graph of hyponym-hypernym relations available in WordNet (Fellbaum 1998; Miller et al. 1990). Moreover, computational tools are available that perform exactly this analysis (e.g., Coh-Metrix: Graesser, McNamara, Louwerse and Cai 2004; McNamara, Louwerse and Graesser 2002). The idea of using an energy function in semantic analysis is also not new, and has been exploited in many variations (including one related to this study: Samsonovich and Ascoli 2010). So, what is new here? The original contribution of the present work is the integration of two ideas, that results in a first-principles-based method of processing hypernym-hyponym relations and a new scale for “abstractness” of words. While the effectiveness of the method seems self-evident from Table 1, its objective evaluation remains to be done elsewhere.

The method is substantially different from well-known techniques, such as the Latent Semantic Analysis and Latent Dirichlet Allocation (Landauer and Dumais, 1997; Landauer et al., 2007). The computed here “abstractness” scale is different from scales computed in those methods, and is probably not reducible to semantic map dimensions in models mentioned above (Osgood et al. 1957; Russell 1980; Bradley and Lang 1999; Gardenfors 2004; Samsonovich and Ascoli 2010), as well as to dimensions of appraisals, affects and emotions (Ortony et al., 1988).

Individual word coordinates calculated in this study are noisy. At the same time, the holistic definition of the “abstractness” scale given by the entire distribution of words is very precise, and allows for further improvement of details of the distribution using standard techniques.

But, does the specific approach presented here truly capture the multifaceted notion of abstractness? The answer is that capturing this multifaceted notion is not a goal pursued in this study, “Abstractness” here is only an intuitive interpretation of the true semantic dimension that is associated with the hyponym-hypernym relations among words. Extracting it numerically is the primary purpose of the presented method. Referring to it as “abstractness” is merely a matter of convenience, or convention. Other words (e.g., “generality”) cold be used instead, or in addition to “abstractness”. The presented method nevertheless captures a definite semantic dimension, which is universal, intuitive, appears to be domain-independent, and is expected to be practically useful. Its precise semantics may not be reducible to one or a few English
words. It is given by the definition of the method itself, and probably cannot be defined more accurately and concisely.

How the method presented in this work can be used in web personalization? One can speculate about possible relations between abstractness and sentiment dimensions, which could make abstractness a useful tool in sentiment analysis. Another idea could be to use abstractness of a pair of words during a matching process that might form the core of a content-based recommender system. These topics need to be considered separately elsewhere.

How many independent universal semantic dimensions can be defined in general? In search for an answer, one may turn to a thread of research that made an attempt to find all semantic primes and came up with the following list (Wierzbicka 1996; Goddard 2002; Goddard and Wierzbicka 2007, 2011): above, after, all, bad, because, before, below, big, body, can, die, do, far, feel, for some time, good, happen, have, hear, here, I, if, inside, kind of, know, like, live, a long time, many/much, maybe, moment, more, move, long, near, not, now, one, other, part of, people/person, the same, say, see, a short time, side, small, some, someone, something/thing, there is, think, this, touch, true, two, very, want, when/time, where/place, word, you. Is “abstractness” defined here reducible to these primitives? This is another question to address elsewhere.

Finally, a practical conclusion in this study is that the WordNet database contains sufficient information to allow one to arrange English words on a line according to their “abstractness”. This information is contained in hyponym-hypernym relations among words. The allocation of words can be achieved by minimization of an energy function defined here in terms of the word coordinates, using the matrix of available hypernym-hyponym relations.

Acknowledgments

I am grateful to Professor Giorgio A. Ascoli for useful discussions of the problem and for his ongoing interest in a related research on word abstractness that goes beyond the present study and will be published elsewhere. I am also grateful to Mr. Thomas Sheehan who processed the data available in WordNet 3.0 and made the matrix of hypernym-hyponym relations available for this study.

References