Proceedings of the 37th Annual Conference of the Cognitive Science Society (2015)

Phrase similarity in humans and machines

Samuel J. Gershman (sjgershm@mit.edu) & **Joshua B. Tenenbaum** (jbt@mit.edu)
Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139, USA

Abstract

Computational models of semantics have emerged as powerful tools for natural language processing. Recent work has developed models to handle compositionality, but these models have typically been evaluated on large, uncontrolled corpora. In this paper, we constructed a controlled set of phrase pairs and collected phrase similarity judgments, revealing novel insights into human semantic representation. None of the computational models that we considered were able to capture the pattern of human judgments. The results of a second experiment, using the same stimuli with a transformational judgment task, support a transformational account of similarity, according to which the similarity between phrases is inversely related to the number of edits required to transform one mental model into another. Taken together, our results indicate that popular models of compositional semantics do not capture important facets of human semantic representation.

Keywords: similarity, semantics, neural networks

Introduction

A central concern of natural language processing is the compact representation of semantic content in words, phrases and documents. Researchers have pursued several different approaches to this problem. One approach is grounded in formal (model-theoretic) semantics, which maps words and sentences onto logical expressions (Montague, 1970). The meaning of a word or sentence, according to this view, is the set of possible worlds in which the corresponding logical expression is true. While general and powerful, this approach has been difficult to apply on a large scale, since the process of mapping arbitrary linguistic fragments to logical expressions is highly non-trivial. A second approach is to build databases of lexical knowledge, like WordNet (Fellbaum, 1998), which offer definitional representations of word meaning. However, this approach does not directly represent the meaning of more complex linguistic units like sentences. A third approach is to derive semantic representations from large corpora by analyzing the co-occurrence of words or phrases in the same context. This approach is grounded in the *distributional hypothesis*: words that occur in similar linguistic contexts have similar meanings (Harris, 1954). This last approach has gained prominence recently, as the combination of massive text data sets and scalable machine learning methods have led to useful applications in core natural language processing tasks, such as information retrieval, sentiment analysis and paraphrase detection.

Our goal in this paper is to present a basic challenge for any computational approach that seeks to capture linguistic meaning at the level of phrases or sentences, or any semantic unit larger than a single word. Our focus is on challenging distributional approaches, where a key open problem is how to capture compositionality: expressing the meaning of a phrase or sentence in terms of the meanings of the words that compose it.

Distributional semantic models are commonly implemented by representing linguistic units as vectors in a high-dimensional space, where spatial proximity encodes semantic relatedness (Turney et al., 2010). If the dimensions of the space are related to features of the linguistic context, then distributional structure will be recapitulated in the spatial structure. A classic example is Latent Semantic Analysis (Landauer & Dumais, 1997), which derives a low-dimensional vector representation from a singular value decomposition of the word-document co-occurrence matrix. In recent years, vector space models have achieved unprecedented success on a number of practical applications, through a combination of larger datasets, more computing power, better learning algorithms, and sophisticated neural network architectures (e.g., Mnih & Hinton, 2009; Collobert et al., 2011).

While most of this work has focused on lexical semantics, a number of researchers have extended vector space models to phrase and sentence meaning (Mitchell & Lapata, 2010; Socher et al., 2011, 2012). Of particular interest is recent work by Socher and colleagues on recursive neural networks, which integrate vector representations with syntactic structure (Socher et al., 2011, 2012). The key idea underlying recursive neural networks is that vector representations for sentences can be generically constructed by recursively applying a composition operator to vectors along a parse tree. Starting at the leaf nodes, the word vectors are composed to form constituent vectors (e.g., noun phrase, verb phrase, etc.), and these vectors are in turn composed until a sentence vector is constructed, corresponding to the root node of the parse tree. By parameterizing the composition operator, and defining an objective function (e.g., classification or reconstruction error), the model can be fit to a text corpus using gradient descent. The choice of composition parameterization, word representation, and objective function is still a matter of active research (Mitchell & Lapata, 2010); in practice, these choices may vary from problem to problem. This is an important development, because it attempts to directly address one of the central criticisms of neural networks—that they lack compositionality, and hence cannot capture the productivity of human thought (Fodor & Pylyshyn, 1988). Although a number of valiant ripostes have been directed against this criticism (Smolensky, 1988; Gelder, 1990), the work of Socher and colleagues was the first

to make neural compositionality work on large text corpora and thereby deliver state-of-the-art performance on paraphrase detection and sentiment analysis.

Recursive neural networks, in common with most distributional semantic models, are typically trained and evaluated on large, diverse text corpora using supervised learning objectives (but see Socher et al., 2011, for an unsupervised objective). While this approach is useful for assessing the overall accuracy of a model, it may not detect subtle failure modes. From a cognitive perspective, we were interested in devising a stimulus set that could be used to evaluate how well state-of-the-art computational models match human semantic representations. To this end, we created a set of simple phrases and used them in a phrase similarity task. The phrases were specially designed to highlight how small changes (in some cases simply changing word order) produce marked changes in similarity judgments. Our stimuli were designed to highlight several of the most basic aspects of compositionality in the structure of noun phrases: (i) the composition of an adjective and a noun, referring to a property and an object respectively, to refer to an object with a specific property, and (ii) the composition of a preposition and two noun phrases, referring to a spatial relation and two objects respectively, to refer to a spatial relation between those two objects.

We found that none of the vector space models that we investigated could adequately capture these patterns of judgment. A further experiment provided support for a transformational theory of similarity, whereby similarity is related to the number of edits required to transform one mental model into another (Hahn et al., 2003; Kemp et al., 2005). We conclude that more theoretical work is needed to design semantic representations that combine the learnability and scalability of current vector-space approaches with a plausible account of compositional meaning at the level of phrases and above, as in more traditional symbolic approaches.

Experiment 1: phrase similarity judgments

Our first experiment collected phrase similarity judgments using a ranking task. We compared several distributional semantic models to human performance on the task.

Methods

Subjects. We recruited 25 human subjects using the Amazon Mechanical Turk web service. All subjects were given informed consent and paid for their participation. The study was approved by the MIT Institutional Review Board.

Procedure. Subjects were shown 30 "base" phrases (e.g., "A young woman in front of an old man") and then asked to rank order 4 transformations of each base phrase in terms of similarity in meaning:

- **Noun change (N)**: "A young man in front of an old woman."
- **Adjective change (A)**: "An old woman in front of a young man."
- **Preposition change (P)**: "A young woman behind an old man."
- **Meaning preservation (M)**: "An old man behind a young woman."

The order of transformations was randomized across trials. Subjects were not shown the transformation labels of these sentences. Table 1 contains the complete set of base sentences.

A young woman in front of an old man.
A black cat in front of a white dog.
A short woman in front of a tall man.
A small bug on a large flower.
A small book on a black table.
A young girl in front of a happy soldier.
A black cow in front of a brown horse.
A sleeping boy in front of a smiling woman.
A white dog on a brown chair.
A happy man in front of an old woman.
A young doctor in front of a smiling patient.
A red apple on green paper.
A white plate on a blue pillow.
A blue pen on a red folder.
A green pear on a brown leaf.
A yellow banana on a green knife.
An orange pepper on a yellow folder.
A plastic bag in front of a brown bottle.
A brown frog on green grass.
A black magazine in front of a white mug.
A pink bowl in front of a blue cup.
A tissue box on yellow paper.
A purple shirt on a green knife.
A young man in front of an angry woman.
A black phone on gray pants.
A rusty bicycle in front of an old fence.
A black marker on a red shirt.
A white sock on black headphones.
An open book in front of a closed window.
A full glass in front of an empty bottle.

Table 1: **Base sentences**

Models We compared 6 computational models to the human data. All models have in common the property that phrase vectors are constructed by recursively applying a composition operator to word vectors. We used the 100-dimensional word vectors from Collobert and colleagues (Collobert et al., 2011), because the recursive neural network models were trained using these

vectors. The word vectors were obtained by training a neural language model to perform a variety of natural language processing tasks, such as part-of-speech tagging and named entity recognition (see Collobert et al., 2011, for details).

The structure of the recursion is determined either by a syntactic parse tree or a simple chain. The models differed in terms of their choice of recursion structure and the choice of elementary composition operator:

- **Sum**: Phrase vector is the sum of the word vectors.
- **Syntactic sum**: Phrase vector is the sum of the word vectors along the parse tree. Specifically, starting at the leaves of the parse tree (corresponding to words in the phrase), the vectors for children of each constituent node are summed and passed through a hyperbolic tangent transformation, and then this sum is passed up the parse tree until the root node vector (corresponding to the entire phrase) is computed. The role of the hyperbolic tangent transformation is to make the phrase representation a nonlinear function of the word representations.
- **Product / Syntactic product**: Same as the Sum / Syntactic sum models, but using elementwise multiplication instead of addition as the elementary composition operator.
- **Recursive autoencoder** (RAE; Socher et al., 2011): Similar to the syntactic sum model, except that the elementary composition operator is parameterized as a linear combination:

$$p = \tanh(W_1 x + W_2 y + b), \tag{1}$$

  where x and y are the children vectors, p is the parent vector, "tanh" is the hyperbolic tangent function, and W_1, W_2 and b are parameters.
- **Matrix-vector recursive neural network** (MV-RNN; Socher et al., 2012): Similar to the RAE, but now constituents are represented by both a vector and a matrix, which allows the model to capture modulatory interactions between constituents:

$$p = \tanh(W_1 Y x + W_2 X y) \tag{2}$$
$$P = W_3 X + W_4 Y, \tag{3}$$

  where X and Y are the children matrices.

For the RAE and the MV-RNN models, we used the parameters that were reported in the original papers (Socher et al., 2011, 2012).

Prediction of similarity rankings We used two different methods to obtain similarity rankings from the vector space models. The first method computed a rank ordering based on the distance between the base phrase vector and each of its transformations. We used Euclidean distance, but essentially indistinguishable results were obtained with cosine and correlation distance (distance functions were obtained by taking the negative of cosine similarity or correlation). The second method used bilinear regression to learn a mapping from phrase vector pairs to dissimilarity. Mathematically, this model has the following form:

$$D(i, j) = x_i \Lambda y_j, \tag{4}$$

where $D(i, j)$ is the dissimilarity between phrases i and j, x_i and y_j are the corresponding vector representations, and Λ is a matrix of regression coefficients. We used leave-one-out cross-validation, fitting the least-squares coefficients to all the phrase pairs except one, and then testing on the held-out pair. The regression model was trained to predict human similarity rankings, which range from 1 (most similar) to 4 (least similar); thus, the model is predicting dissimilarity.

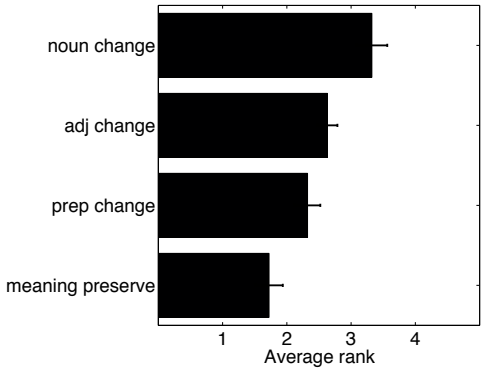

Figure 1: **Experiment 1 results**. 1 = most similar, 4 = least similar. Error bars represent standard error of the mean.

Results and discussion

Humans show a systematic pattern in their similarity rankings (Figure 1): the meaning-preserving transformation is judged most similar, followed by preposition change, adjective change, and finally noun change. To quantify this pattern, we computed pairwise t-tests between the rankings of the different transformations; all tests were statistically significant ($p < 0.05$) except for the difference between the meaning preserving change and the preposition change.

The model predictions, using the vector distance method, are shown in Figure 2. These results demonstrate that none of the models described above can adequately capture the behavioral results. None of the models correctly predicts the rank ordering exhibited by humans. A similar conclusion can be drawn from the model predictions using the bilinear method (Figure 3).

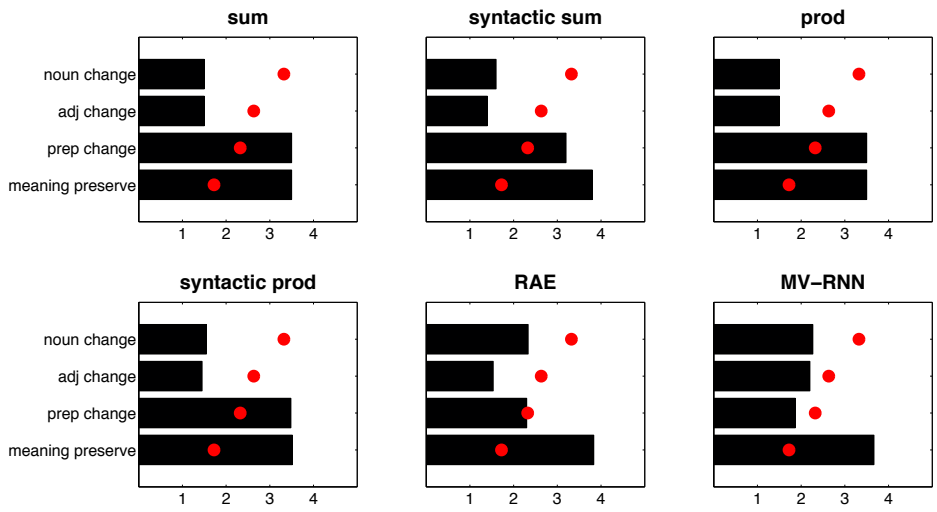

Figure 2: **Model predictions, vector distance method**. Human behavioral data are superimposed in red.

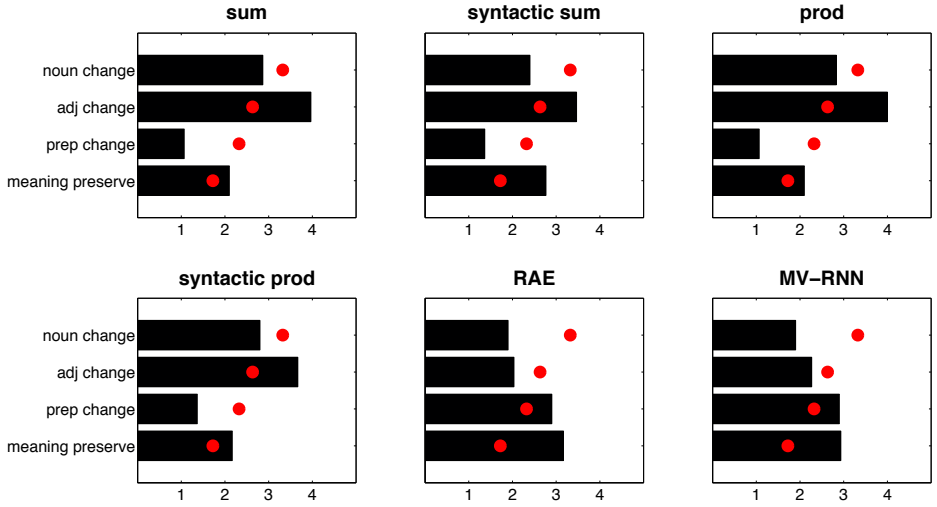

Figure 3: **Model predictions, bilinear regression method**. Human behavioral data are superimposed in red.

To quantify the fit between models and human judgment, we computed correlations for both the vector distance and bilinear regression methods. The results are shown in Figure 4. In most cases, the correlations are not significantly different from 0. In the case of the sum and product models, the correlation is significantly greater than 0 ($p < 0.05$). This is largely driven by their correct prediction that preposition and meaning preserving changes are more similar than noun and adjective changes. However, visual inspection of Figure 3 reveals that they incorrectly order preposition vs. meaning preserving change, as well as noun vs. adjective change. Thus, the significant correlation is not a particularly resounding endorsement of their descriptive adequacy.

One direction for future research is to explore alternative regression models. For example, instead of learning a bilinear model, one could learn a model that is linear in the elementwise difference or ratio of the two vectors. One could also explore using regularization or placing constraints on the bilinear model (e.g., requiring that Λ be low rank).

Experiment 2: event transformation judgments

One possible interpretation of the results from Experiment 1 is that they reflect a process of mental model building (Johnson-Laird, 1983; Tenenbaum et al., 2011). In particular, semantic similarity may reflect the difficulty of modifying the mental model of the base phrase to align with the transformed phrase. In the simplest terms, we can think of each of our base sentences as establishing a mental model with three constituents expressing the attributes of and relations among entities in a scene. For instance, the mental model of "A young

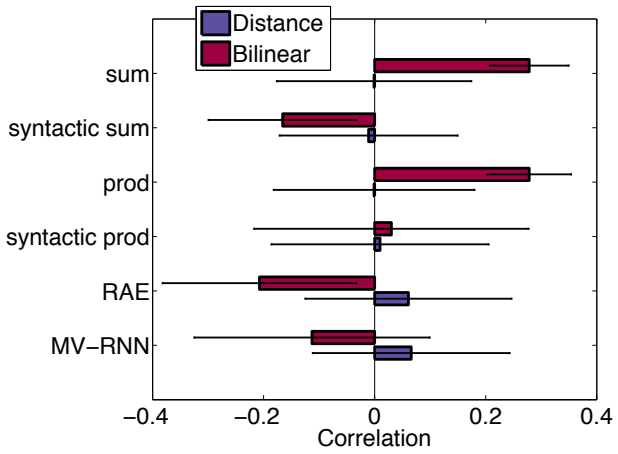

Figure 4: **Correlation between model predictions and human similarity judgments**. Error bars represent bootstrapped 95% confidence intervals.

woman in front of an old man" would consist of these constituents: (*i*) A woman in front of a man, (*ii*) a young woman, and (*iii*) an old man. The different transformations of this base sentence can be ordered in terms of how many semantic constituents are changed: M (0 changes), P (changes *i*), A (changes *ii* and *iii*), N (changes all three). This account of similarity judgment has much in common with the representational distortion account of similarity (Hahn et al., 2003), according to which the similarity between two entities is the complexity of the operations required to distort one into the other (see Kemp et al., 2005, for a Bayesian treatment of this idea). We believe that developing models of compositional semantics along these lines is a promising avenue of future research. To explore this idea, we slightly modified the procedure from Experiment 1 to elicit judgments of transition probability rather than similarity.

Methods

Subjects. We recruited 20 human subjects using the Amazon Mechanical Turk web service. All subjects were given informed consent and paid for their participation. The study was approved by the MIT Institutional Review Board.

Procedure. The procedure and stimuli were identical to the procedure for Experiment 1, except for a change of instructions. Instead of asking subjects to rank phrases according to their similarity in meaning, we asked them to imagine the phrases as scenes, and rank the transformations according to the likelihood that each transformed scene would occur following the base scene (i.e., the transition probability). In other words, we asked subjects to rank phrases according to their transition probability. We refer to this as the *transformation ranking*.

Results and discussion

Figure 5 (left) shows the transformation rankings reported by our subjects. These results closely resemble the phrase similarity rankings reported by subjects in Experiment 1. To establish this correspondence at an item level, we plotted the transformation rankings against the similarity rankings for each phrase (right panel of Figure 5). The two rankings are strongly correlated ($r = 0.87, p < 0.00001$). These results support the transformational view of similarity (Hahn et al., 2003; Kemp et al., 2005), which holds that similarity judgments reflect the number (and probability) of transformations required to transform one object into another. In our case the "objects" are descriptions of scenes, and the transformations are events that cause scenes to change.

General discussion

Our experimental results are deflationary for several prominent models of compositional semantics. While the models are effective at capturing some aspects of semantics in large corpora, they decisively fail in carefully constructed test cases such as the ones presented here (see also Pham et al., 2013). Some of this may be attributable to the fact that the models were not trained to perform the task given to subjects, but the poor performance of the bilinear model (which is trained to perform the task) suggests that other methods may be required to match human performance. There may also be other, deeper problems with vector-based compositional semantics, which are hard to diagnose because their semantic information content and algebraic composition operators are opaque to interpretation.

We have suggested that a transformational account (Hahn et al., 2003; Kemp et al., 2005) may provide the basis for a better theory of phrase similarity. Evidence in support of this account was provided by Experiment 2, which showed that similarity judgments can be predicted by transformation judgments. Models based on recurrent neural networks (Sutskever et al., 2011) might be able to capture certain kinds of transformations; however, we believe that doing this appropriately for natural language requires a theory of how syntactic structure relates to event knowledge. It is presently unclear whether such a theory can be obtained via a purely distributional approach.

Our findings pose a general challenge to computational models of phrase similarity. We hope that the failures of the models explored in this paper will stimulate the development of new approaches, perhaps based on some combination of probabilistic world models and vector space representations.

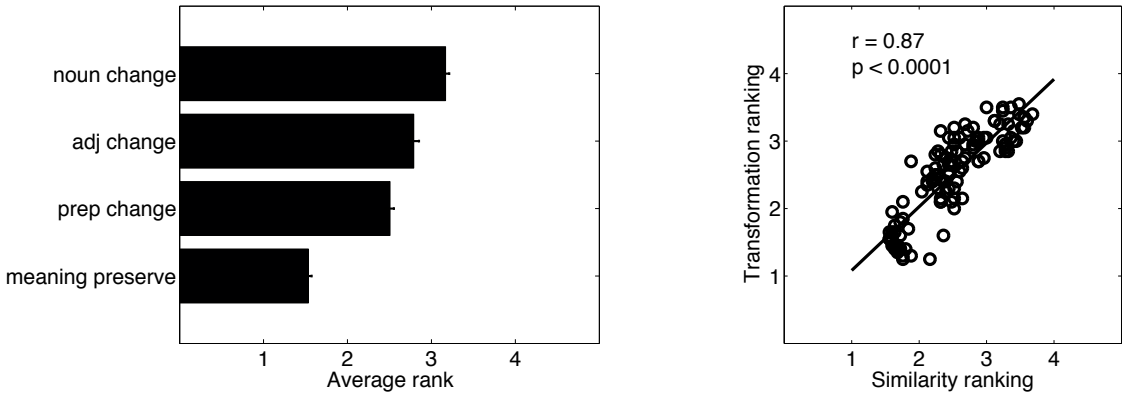

Figure 5: **Experiment 2 results**. (Left) Transformation rankings. Error bars represent standard error of the mean. (Right) Transformation rankings plotted against similarity rankings. Each point represents a single phrase.

Acknowledgments

We are grateful to Sam Ritter, Richard Socher, Tim O'Donnell and Anatole Gershman for helpful discussions. This work was supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Air Force Research Laboratory (AFRL), under contract FA8650-14-C-7358. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, AFRL, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, *12*, 2493–2537.

Fellbaum, C. (1998). *WordNet*. Wiley Online Library.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3–71.

Gelder, T. (1990). Compositionality: A connectionist variation on a classical theme. *Cognitive Science*, *14*, 355–384.

Hahn, U., Chater, N., & Richardson, L. B. (2003). Similarity as transformation. *Cognition*, *87*, 1–32.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*, 146–162.

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press.

Kemp, C., Bernstein, A., & Tenenbaum, J. B. (2005). A generative theory of similarity. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, (pp. 1132–1137). Citeseer.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, *34*, 1388–1429.

Mnih, A., & Hinton, G. E. (2009). A scalable hierarchical distributed language model. In *Advances in Neural Information Processing Systems*, (pp. 1081–1088).

Montague, R. (1970). *English as a Formal Language*. Ed. di Comunità.

Pham, N., Bernardi, R., Zhang, Y. Z., & Baroni, M. (2013). Sentence paraphrase detection: When determiners and word order make the difference. In *Proceedings of the IWCS 2013 Workshop: Towards a Formal Distributional Semantics*.

Smolensky, P. (1988). The constituent structure of connectionist mental states: A reply to fodor and pylyshyn. *The Southern Journal of Philosophy*, *26*, 137–161.

Socher, R., Huang, E. H., Pennin, J., Manning, C. D., & Ng, A. Y. (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in Neural Information Processing Systems*, (pp. 801–809).

Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, (pp. 1201–1211). Association for Computational Linguistics.

Sutskever, I., Martens, J., & Hinton, G. E. (2011). Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning*, (pp. 1017–1024).

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*, 1279–1285.

Turney, P. D., Pantel, P., et al. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.