Hierarchical Reasoning with Distributed Vector Representations

Cody Kommers (cydeko@ucla.edu)
Department of Psychology, University of California, Los Angeles
Los Angeles, CA 90095 USA

Volkan Ustun (ustun@ict.usc.edu)
Institute for Creative Technologies
Playa Vista, CA 90094 USA

Abram Demski (demski@usc.edu)
Department of Computer Science, University of Southern California
Institute for Creative Technologies
Playa Vista, CA 90094 USA

Paul Rosenbloom (rosenbloom@usc.edu)
Department of Computer Science, University of Southern California
Institute for Creative Technologies
Playa Vista, CA 90094 USA

Abstract

We demonstrate that distributed vector representations are capable of hierarchical reasoning by summing sets of vectors representing hyponyms (subordinate concepts) to yield a vector that resembles the associated hypernym (superordinate concept). These distributed vector representations constitute a potentially neurally plausible model while demonstrating a high level of performance in many different cognitive tasks. Experiments were run using DVRS, a word embedding system designed for the Sigma cognitive architecture, and Word2Vec, a state-of-the-art word embedding system. These results contribute to a growing body of work demonstrating the various tasks on which distributed vector representations perform competently.

Keywords: hierarchical reasoning; word embeddings; language modeling; concepts; distributed representations

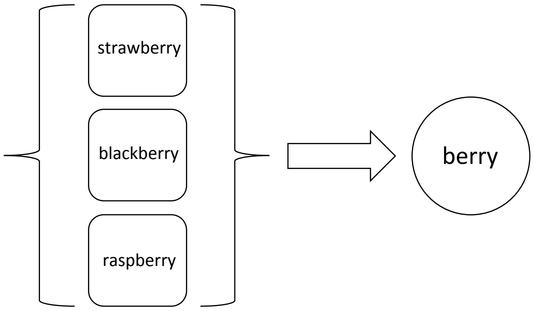

Figure 1: The normalized summation of the vectors representing hyponyms *strawberry*, *blackberry*, and *raspberry* yields a vector resembling the hypernym *berry*.

Introduction

In this paper, we demonstrate that distributed vector representations are capable of performing hierarchical reasoning by inferring the appropriate category from a set of category members (Figure 1). This capability is one among many varieties of tasks that recent methods for learning distributed vector representations, also known as word embeddings, have been shown to perform competently. These capabilities include: language modeling (Bengio et al., 2006; Mikolov, 2012), natural language understanding (Collobert & Weston, 2008; Zhila et al., 2013), machine translation (Mikolov et al., 2013a; Zou et al., 2013), image labeling (Frome et al., 2013), paragraph representation (Le & Mikolov, 2014), and relational extraction (Socher et al., 2013).

In addition to state-of-the-art performance, one major advantage of these vector models is their supposed neural-plausibility (Blouw & Eliasmith, 2013). At a gross level of abstraction, concepts are represented in the brain as distributed networks of neural activation throughout cortical and subcortical regions (Rissman & Wagner, 2011). Distributed vector representations attempt to approximate these distributed networks of activation. Intuitively, if the semantic discrepancies between two concepts, such as “dog” and “cat”, can successfully be encoded as distinct patterns of neural firings, then it follows that the discrepancies could also be encoded as distinct patterns of values in a vector (Hinton, 1984). Furthermore, Kelly and West (2012) argue that vector representations constitute both symbolic and subsymbolic representation, allowing distributed vector representations to provide a comprehensive analysis of a cognitive process.

An example of the utility of distributed vector representations as a model of cognitive phenomena is their application in analogical reasoning. The use of vectors to model analogical reasoning dates back to Rumelhart and Abrahamson (1973). More recent examples of distributed vector representations for performing analogy are presented in Zhilia et al. (2013) and Socher et al. (2013). In these examples, analogical reasoning with distributed vector representations can be performed by vector arithmetic. For example, the difference between vectors representing *woman* and *man* is approximately the same as the difference between vectors representing *queen* and *king*:

$$[W] - [M] \approx [Q] - [K] \quad (1)$$

This equation can also be interpreted as an analogy, *woman:man::queen:king*.

Analogical reasoning with distributed vector representations can be incorporated into existing cognitive models. Upon reorganization, the equation for analogical reasoning with vectors (Equation 2) successfully maps onto an excerpt of the cognitive model of analogical reasoning presented in Holyoak (2012). $[W]-[M]$ maps onto the source; $+[K]$ maps onto the target; and $\approx [Q]$ maps onto the inference.

$$[W] - [M] + [K] \approx [Q] \quad (2)$$

In this paper, we demonstrate that distributed vector representations can perform hierarchical reasoning in a similar manner to how they perform analogical reasoning (i.e., with simple vector arithmetic). The present demonstration contributes to accumulating work exemplifying the capabilities of distributed vector representations. Additionally, it is unclear to what extent distributed vector representations are capable of modeling human cognition, and hierarchical reasoning, like analogical reasoning, is among the features for which a comprehensive model of human cognition must account. Thus, the present results lend support for the ability of distributed vector representations to model human cognition.

It has previously been hypothesized that distributed vector representations can accurately represent hierarchical information (e.g., Lenci & Benotto, 2012; Erk, 2009a; Erk, 2009b; McDonald & Ramscar, 2001; Geffet & Dagan, 2005). Lenci & Benotto (2012) define this as the *distributional inclusion hypothesis*: “if *u* is a semantically narrower term than *v*, then a significant number of salient distributional features of *u* is included in the feature vector of *v* as well.” Previous experiments have attempted to quantify the phenomenon of vectors representing hyponyms sharing a common set of characteristic features of the associated hypernym.

A representative example of these previous experiments is Geffet and Dagan (2005), which developed an automated word-level feature inclusion testing method, called the Inclusion Testing Algorithm (ITA). For each pair of vectors representing a hypernym and a hyponym, ITA computes a set of characteristic features for the hypernym vector and tests if those features are also included in the vector of the hyponym. This inclusion occurred in 86% percent of their tested pairs. In other words, the vector that represents a category member contains the information that characterizes it as a member of the associated category.

The results presented in this paper derive from the same hypothesis as the above results, but further the empirical analysis. Specifically, instead of comparing a single hyponym vector with a single hypernym vector, we compare sets of hyponym vectors with the common hypernym. This is an advance for cognitive modeling of hierarchical reasoning because it transitions from purely representational (i.e., that vectors can represent the characteristics) to algorithmic (i.e., deriving the hypernym shared among hyponyms).

The results in this paper are derived chiefly from distributed vector representations learned by DVRS (Ustun et al., 2014), a word embedding system designed for the Sigma cognitive architecture (Rosenbloom, 2013). DVRS learns real-valued lexical (meaning) vectors in an unsupervised manner from large, shallow information sources based chiefly on co-occurrence and skip-gram algorithms. DVRS is intended to be implemented within the Sigma cognitive architecture and thus strives to maximize performance while retaining its integrity as a cognitive model. As a point of comparison, results are also presented from vectors learned by Word2Vec (Mikolov et al., 2013b), a state-of-the-art word embedding system.

DVRS draws inspiration from BEAGLE (Jones & Mewhort, 2007), but relies on skip-grams rather than n-grams and replaces circular convolution with pointwise multiplication. Representations learned by BEAGLE have been hypothesized to encode hierarchical information. This is suggested by the tendency of the representations to cluster hierarchically (e.g., vehicles with other vehicles and birds with other birds), but no formal demonstration exists, to our knowledge, in the capacity demonstrated in this paper.

To obtain the present results, vectors of 200 dimensions were trained for both DVRS and Word2Vec with their respective default settings on the first 10^9 bytes of a Wikipedia dump from March 3, 2006 (enwik9)[1]. The data were preprocessed to convert all text to lower case, convert numbers to text, and eliminate links and other references[2].

Experiments

Four experiments were run on three different corpora of hypernym-hyponym sets to demonstrate hierarchical reasoning with distributed vector representations. The first three experiments measure aptitude for hierarchical reasoning with distributed vector representations. The fourth experiment measures the number of neighbors for both hyponyms and their associated hypernyms to evaluate the hypothesis that more general concepts (i.e., hypernyms) have more neighbors.

In each of the first three experiments, a set of N vectors representing hyponyms were summed; the result was normalized and the closest M vectors in the lexicon, as measured by cosine similarity, were considered as potential hypernyms. If the appropriate hypernym was among the M closest vectors, then the trial was counted as correct. As with analogical reasoning, the vector calculation for hierarchical reasoning can be expressed by an equation of vector arithmetic:

$$\frac{\sum_1^N [h_n]}{|\sum_1^N [h_n]|} \approx [h_{category}] \quad (3)$$

[1] Obtained from http://cs.fit.edu/mmahoney/compression/textdata.html.

[2] Script provided by Matt Mahoney with the Wikipedia dump.

N is the number of hyponyms being summed, $[h_n]$ is the vector representing the hyponym, and $[h_{category}]$ is the vector representing the hypernym. For example, the normalized summation of the vectors for hyponyms *strawberry*, *blackberry*, and *raspberry* yields a vector resembling the hypernym *berry* (Figure 1). In other words, the vectors can be used to judge that *strawberry*, *blackberry*, and *raspberry* are members of the category *berry*.

The use of three different corpora show the robustness of this effect, independent of any idiosyncrasy of the data. In addition to using different corpora, three values were varied to show a range of effectiveness in performing the task: (1) the word embedding system under consideration (labeled *System* in results tables), (2) N, the number of hyponyms to sum (labeled *#Hypo*), and (3) M, the number of closest vectors in the lexicon that will be considered as potential hypernyms (labeled *#Hyper*). The table shows the number correct out of the total possible, which varies based on how the categories were grouped.

Results are shown for both DVRS and Word2Vec in most cases. DVRS is the focus of these results, because it is most concerned with explanatory power as a cognitive model. Word2Vec is shown as a point of comparison, because it is a state-of-the-art word embedding system. Results from Word2Vec are shown mostly for the trials in which the best performance is expected (i.e., *#Hypo*=10). This allows for comparison of performance between DVRS and Word2Vec, while elaborating on several more detailed cases with DVRS (i.e., cases with fewer summed hyponyms).

The number of hyponyms was varied to demonstrate that, in principle, it is possible to derive hypernyms from a relatively small set of hyponyms. There are many anecdotal cases in which the correct hypernym can be derived from only two hyponyms. Accordingly, even though performance is weaker with sets of fewer hyponyms, there are still examples of successful trials.

The number of closest vectors considered as potential hypernyms is varied to demonstrate that, even if the targeted hypernym vector is not the top result, it is among the top results. That is, the vector resulting from the summed hyponyms consistently resembles the vector of the hypernym, even though it may not be the best match.

Experiment 1

The first data set is from McRae et al. (2005), a corpus consisting of human-generated data on semantic feature norms for 541 basic-level concepts. Seven hundred and twenty-five participants were recruited to label semantic features for the concepts. Each concept was labeled with 10 unprompted semantic features by at least 30 of the subjects. A semantic feature was included as a norm if more than three participants included the feature for the same concept. Hypernymy was among the semantic features coded by the researchers. A subset of 535 concepts were used in the present experiment; this subset consisted of basic-level concepts for which at least three shared a hypernym.

The number of hyponyms per hypernym varied from three (in the case of six hypernyms) to 98 (in the case of one hypernym, *animal*). For the trials labeled *All* in the *#Hypo* column, every hyponym was summed, regardless of number. For trials with 3 or 10 hyponyms, hyponyms were separated into appropriately sized subsets. For example, in a trial with 10 summed hyponyms, the hyponyms of animal (with 98 total hyponyms) were separated into nine different subsets. These sets were created by alphabetical ordering (e.g., the top 10 in alphabetical order constituted the first set) and remaining hyponyms were not included in the test.

For the present purposes, this data set contains a small amount of noise. These data were generated by labeling semantic features that were judged to be associated with basic-level concepts. That is to say, these features are not necessarily concerned with selecting the canonical hypernym of the hyponyms, merely a hypernym that is accurate. It does not necessarily follow that the labeled hypernym is the best hypernym that follows from the associated hyponyms. Thus, the distributed vector representations may produce an answer that is acceptable, but merely not listed in these data. While this ambiguity does not nullify the utility of these data, it presents an issue that may negatively affect the performance.

Results are shown in Table 1. DVRS performs best, as expected, in the case of summing 10 hyponyms and considering 10 hypernyms; at 44% the performance is consistent, though not tremendous. DVRS significantly outperforms Word2Vec in every case. It is unclear exactly why this discrepancy in performance occurred.

Table 1: Hierarchical reasoning results on McRae et al. (2005) data.

System	#Hypo	#Hyper	Corr.	Total	Acc.
DVRS	All	1	9	39	23.1%
DVRS	3	1	18	167	10.8%
DVRS	3	10	54	167	32.3%
DVRS	10	1	8	34	23.5%
DVRS	10	10	15	34	**44.1%**
Word2Vec	All	1	3	39	7.7%
Word2Vec	10	1	1	34	2.9%
Word2Vec	10	10	5	34	**14.7%**

Experiment 2

The second data set is derived from WordNet (Miller, 2005), a standard database of semantic relationships between words. One hundred and forty-seven basic-level concepts were chosen as hypernyms for which WordNet supplied hyponyms. Though WordNet supplied the hyponyms, these hypernyms were chosen by the authors under no systematic criterion. For that reason, this set of basic-level concepts cannot be claimed to achieve the same standard of empirical disinterest as the McRae et al. (2005) data set. However, there appears to be no qualitative difference between the kinds of basic-level concepts that appear in these data versus those from the data in McRae et al. (2005).

WordNet choices for hyponyms, while empirical, contain significant noise. That is to say, in many cases they do not represent what could be considered canonical categorizations of hypernymy as one could imagine might be judged by a human. For example, *schizocarp, pyxis, rowanberry,* and *drupe* appear as hyponyms for hypernym *fruit*; *shamanism, zoroastrianism, mithraism,* and *hindooism* [*sic*] appear as hyponyms for hypernym *religion*; *thanatology, cryptanalysis, agrobiology,* and *architectonics* appear as hyponyms for hypernym *science*. While these categorizations may not be inaccurate, they do not constitute the most representative set of human judgments of hypernymy.

Results are shown in Table 2. As with Experiment 1, DVRS outperforms Word2Vec in every case, though by a smaller margin. In the best-performance-expected trial (10 hyponyms summed and 10 hypernyms considered), DVRS obtains 50.0% accuracy, which is comparable to its performance of 44.1% under the same criteria in Experiment 1. In contrast, Word2Vec more than doubles its accuracy for the same best-performance-expected criteria between Experiment 1 (14.7%) and Experiment 2 (35.4%). Results of hyponym summations begin with sets of five (instead of three) to adjust for noise associated with WordNet hyponyms.

Table 2: Hierarchical reasoning results on WordNet data.

System	#Hypo	#Hyper	Corr.	Total	Acc.
DVRS	5	1	24	119	20.2%
DVRS	5	5	36	119	30.3%
DVRS	5	10	55	119	46.2%
DVRS	10	1	19	82	23.2%
DVRS	10	5	27	82	32.9%
DVRS	10	10	41	82	**50.0%**
Word2Vec	10	1	10	82	12.2%
Word2Vec	10	5	22	82	26.8%
Word2Vec	10	10	29	82	**35.4%**

Experiment 3

The third data set consists of 58 sets of eight subordinate concepts selected by the authors to constitute a data set that would be most likely to result in correct answers from DVRS. While these data do not represent a randomly selected sample, they are judged by the authors to be a data set without the ambiguity of the data from McRae et al. (2005) or the noise of the data from WordNet. There was no systematic criterion by which these data were selected; they were selected only by if the authors thought the system should be capable of producing a category shared by all members. Thus, they should be interpreted as an upper bound of the capabilities of hierarchical reasoning with distributed vector representations in the present paradigm; that is to say, these are the results of a hand-selected data set and a charitable judging criterion.

For each set of *N* hyponyms, the result from the closest *M* vectors was judged to be correct if it represented any commonality between the hyponyms. This could include a common category, a common entity of which all vectors are members, or a common trait. For example, a trial was considered correct if hyponyms *Monterrey*, *Bakersfield*, and *Riverside* yield a hypernym such as *city*, the common entity of which all are members such as *California*, or a common attribute such as *Californian*. A trial was considered incorrect if the summation yielded results such as similar category members (e.g., *Merced*) or wholly unrelated concepts.

Results are shown in Table 3. While both systems demonstrated their best respective results, there were still failed instances by both. These failed instances most likely do not come from a lack of examples by which a sufficient representation can be learned, but solely a failure to encode hierarchical information. For example, with the *California* example mentioned above, DVRS got the trial incorrect because the hierarchical information was insufficiently represented, not necessarily because there were too few encounters with the associated words. This claim is corroborated by a correct response from Word2Vec in the *California* case.

Qualitative analysis suggests that often when the answer is completely incorrect, the result is a another hyponym instead of a hypernym (i.e., a category member rather than the category). For example, a set of vectors representing cardinal directions including *north*, *west*, *northwest*, etc. yields *southeastern* rather than *directions* or *cardinal*. This seemed to be the case for Experiments 1 and 2 as well.

Table 3: Hierarchical reasoning results on data selected by authors. These results may be considered an upper-bound on performance for hierarchical reasoning with the present word embedding systems.

System	#Hypo	#Hyper	Corr.	Total	Acc.
DVRS	8	1	28	58	48.3%
DVRS	8	10	50	58	**86.2%**
Word2Vec	8	1	6	58	10.4%
Word2Vec	8	10	36	58	**62.1**%

Experiment 4

A subset of the WordNet hypernym-hyponym sets used for testing hierarchical reasoning in Experiment 2 were used to compare the number of neighbors between hypernyms and hyponyms in the associated vector space. This subset consisted of 39 hypernym-hyponym sets which DVRS got correct in Experiment 2 (i.e., hierarchical information was successfully encoded) and 39 hypernym-hyponym sets which DVRS did not get correct in Experiment 2 (i.e., hierarchical information was not successfully encoded). We hypothesized that, in comparison between a category and a subcategory, the more general concept would have more neighbors (Figure 3). Additionally, for sets in which the relevant hierarchical information was shown to be successfully encoded in Experiment 2, the effect should be more consistent than those sets that did not demonstrate success in Experiment 2. The present results support this hypothesis.

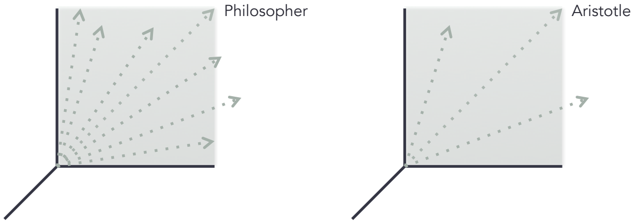

Figure 2: In this example, *Philosopher* is more general than *Aristotle*; thus, its vector would be hypothesized to have more neighbors.

For each trial, the number of neighbors within a cosine similarity of 0.7 was compared between the hypernym and the average of three corresponding hyponyms. The three hyponyms were randomly selected from the set of all WordNet hyponyms corresponding to the hypernym. If the number of neighbors was higher for the hypernym than the average hyponym, then the trial was counted as correct.

Results are shown in Table 4. As expected, vectors representing hypernyms consistently have more neighbors within a cosine similarity of 0.7 than their hyponyms. In the case of hypernym-hyponym sets where the hierarchical relationship is demonstrated to be encoded (i.e., correct in Experiment 2), this effect was seen in 89.7% of tested cases. If the hierarchical relationship was demonstrated to not be successfully encoded (i.e., incorrect in Experiment 2), this effect is closer to chance (66.7%). According to our hypothesis, if the hierarchical information is not encoded, then the probability of the hypernym having more neighbors should be the same as in a comparison with any other word (i.e., 50%). Thus, it appears that some hierarchical information is encoded in those cases that are unsuccessful in Experiment 2, but not sufficient for robust performance.

Table 4: Vectors representing hypernyms consistently have more neighbors within a cosine similarity of 0.7.

System	Hierarchy encoded?	Corr.	Total	Acc.
DVRS	Yes	35	39	**89.7%**
DVRS	No	26	39	66.7%

Discussion

The present experiments demonstrate that distributed vector representations can successfully encode hierarchical information. The discrepancy in performance between DVRS and Word2Vec suggests that not all methods of learning such vectors yield equally successful representations. Additionally, an intriguing relationship has been uncovered between concept generality and the number of neighbors in the associated vector space.

Proposed Geometric Intuition

What is the intuition behind how distributed vector representations are capable of representing this hierarchical information? One plausible explanation is that this effect is merely a function of word frequency. In this case, *Philosopher* has more neighbors than *Aristotle* simply because it is more frequent in the corpus on which the vectors were learned. Another explanation which may also contribute is that more general concepts take on properties of a hyperplane on which subordinate concepts lay (Figure 2).

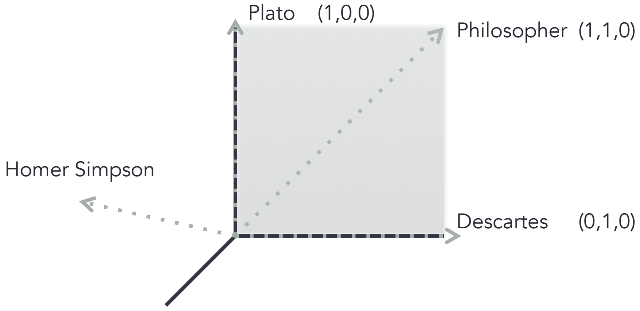

Figure 3: More general concepts may take on features of a hyperplane on which the associated subordinate concepts lay.

In the 3-dimensional case depicted in Figure 3, the hypernym *Philosopher* encodes relevant feature information related to its hyponyms *Plato* and *Descartes*. *Philosopher* may be thought of as having hyperplane-like properties because its vector is yielded by the summation of its hyponyms. Related concepts, such as *Plato* and *Descartes*, lay on the hyperplane created by *Philosophy* while unrelated concepts, such as *Homer Simpson*, do not. While in three dimensions this is an implausible scenario for a complex ontology of concepts, it is plausible for a high dimensional space (e.g., 200), as with distributed vector representations.

This interpretation appears to be in line with the aforementioned distributional inclusion hypothesis, in which subordinate concepts include features of their superordinate concepts. More results will be necessitated to explore this connection.

Acknowledgments

This work has been sponsored by the U.S. Army. Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government. No official endorsement should be inferred.

References

Barbu, E. (2009). Acquisition of common sense knowledge for basic level concepts. *RANLP*, 23–27.

Bengio, Y., Schwenk, H., Senécal, J.-S., Morin, F., & Gauvain, J.-L. (2006). Neural probabilistic language models. *Innovations in Machine Learning*, 137–186.

Blouw, P., & Eliasmith, C. (2013). A neurally plausible encoding of word order information into a semantic vector space. *35th Annual Conference of the Cognitive Science Society*, 1905–1910.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. *Proceedings of the 25th international conference on Machine learning*, 160–167.

Eliasmith, C., & Thagard, P. (2001). Integrating structure and meaning: A distributed model of analogical mapping. *Cognitive Science*, *25*(2), 245–286.

Erk, K. (2009a). Representing words as regions in vector space. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, 57–65.

Erk, K. (2009b). Supporting inferences in semantic space: representing words as regions. *Proceedings of the Eighth International Conference on Computational Semantics*, 104–115.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al. (2013). Devise: A deep visual-semantic embedding model. *Advances in Neural Information Processing Systems*, 2121–2129.

Geffet, M., & Dagan, I. (2005). The distributional inclusion hypotheses and lexical entailment. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 107–114.

Hinton, G. (1984). *Distributed representations* (Tech. Rep. No. CMU-CS-84-157). Pittsburgh, PA: Carnegie Mellon University, School of Computer Science.

Holyoak, K. J. (2012). Analogy and relational reasoning. *The Oxford Handbook of Thinking and Reasoning*, 234–259.

Jones, M. N., & Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological review*, *114*(1), 1.

Kelly, M. A., & West, R. L. (2012). From vectors to symbols to cognition: The symbolic and sub-symbolic aspects of vector-symbolic cognitive models.

Kiefer, M., & Pulvermüller, F. (2012). Conceptual representations in mind and brain: theoretical developments, current evidence and future directions. *Cortex*, *48*(7), 805–825.

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

Lenci, A., & Benotto, G. (2012). Identifying hypernyms in distributional semantic spaces. *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, 75–79.

Levy, S. D., & Gayler, R. (2008). Vector symbolic architectures: A new building material for artificial general intelligence. *Proceedings of the 2008 Conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, 414–418.

McDonald, S., & Ramscar, M. (2001). Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. *Proceedings of the 23rd annual conference of the cognitive science society*, 611-616.

McRae, K., Cree, G., Seidenberg, M., & Mcnorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547-559.

Mikolov, T. (2012). Statistical language models based on neural networks. *Presentation at Google, Mountain View, 2nd April.*

Mikolov, T., Le, Q. V., & Sutskever, I. (2013a). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26*, 3111–3119.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, *34*(8), 1388–1429.

Rissman, J., & Wagner, A. D. (2012). Distributed representations in memory: insights from functional brain imaging. *Annual Review of Psychology*, *63*, 101–128.

Rosenbloom, P. S. (2013). The sigma cognitive architecture and system. *AISB Quarterly*, *136*, 4–13.

Rumelhart, D. E., & Abrahamson, A. A. (1973). A model for analogical reasoning. *Cognitive Psychology*, *5*(1), 1–28.

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620.

Socher, R., Chen, D., Manning, C. D., & Ng, A. (2013). Reasoning with neural tensor networks for knowledge base completion. *Advances in Neural Information Processing Systems*, 926–934.

Ustun, V., Rosenbloom, P. S., Sagae, K., & Demski, A. (2014). Distributed vector representations of words in the sigma cognitive architecture. *Artificial General Intelligence*, 196–207.

Van Der Vet, P. E., & Mars, N. J. (1998). Bottom-up construction of ontologies. *Knowledge and Data Engineering, IEEE Transactions on*, *10*(4), 513–526.

Zhila, A., Yih, W., Meek, C., Zweig, G., & Mikolov, T. (2013). Combining heterogeneous models for measuring relational similarity. *HLT-NAACL*, 1000–1009.

Zou, W. Y., Socher, R., Cer, D. M., & Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation. *EMNLP*, 1393–1398.