Word Embeddings as Metric Recovery in Semantic Spaces

Tatsunori B. Hashimoto, David Alvarez-Melis and **Tommi S. Jaakkola**
CSAIL, Massachusetts Institute of Technology
{thashim, davidam, tommi}@csail.mit.edu

Abstract

Continuous word representations have been remarkably useful across NLP tasks but remain poorly understood. We ground word embeddings in semantic spaces studied in the cognitive-psychometric literature, taking these spaces as the primary objects to recover. To this end, we relate log co-occurrences of words in large corpora to semantic similarity assessments and show that co-occurrences are indeed consistent with an Euclidean semantic space hypothesis. Framing word embedding as *metric recovery* of a semantic space unifies existing word embedding algorithms, ties them to manifold learning, and demonstrates that existing algorithms are consistent metric recovery methods given co-occurrence counts from random walks. Furthermore, we propose a simple, principled, direct metric recovery algorithm that performs on par with the state-of-the-art word embedding and manifold learning methods. Finally, we complement recent focus on analogies by constructing two new inductive reasoning datasets—series completion and classification—and demonstrate that word embeddings can be used to solve them as well.

1 Introduction

Continuous space models of words, objects, and signals have become ubiquitous tools for learning rich representations of data, from natural language processing to computer vision. Specifically, there has been particular interest in word embeddings, largely due to their intriguing semantic properties (Mikolov et al., 2013b) and their success as features for downstream natural language processing tasks, such as named entity recognition (Turian et al., 2010) and parsing (Socher et al., 2013).

The empirical success of word embeddings has prompted a parallel body of work that seeks to better understand their properties, associated estimation algorithms, and explore possible revisions. Recently, Levy and Goldberg (2014a) showed that linear linguistic regularities first observed with `word2vec` extend to other embedding methods. In particular, explicit representations of words in terms of co-occurrence counts can be used to solve analogies in the same way. In terms of algorithms, Levy and Goldberg (2014b) demonstrated that the global minimum of the skip-gram method with negative sampling of Mikolov et al. (2013b) implicitly factorizes a shifted version of the pointwise mutual information (PMI) matrix of word-context pairs. Arora et al. (2015) explored links between random walks and word embeddings, relating them to contextual (probability ratio) analogies, under specific (isotropic) assumptions about word vectors.

In this work, we take *semantic spaces* studied in the cognitive-psychometric literature as the prototypical objects that word embedding algorithms estimate. Semantic spaces are vector spaces over concepts where Euclidean distances between points are assumed to indicate semantic similarities. We link such semantic spaces to word co-occurrences through semantic similarity assessments, and demonstrate that the observed co-occurrence counts indeed possess statistical properties that are consistent with an underlying Euclidean space where distances are linked to semantic similarity.

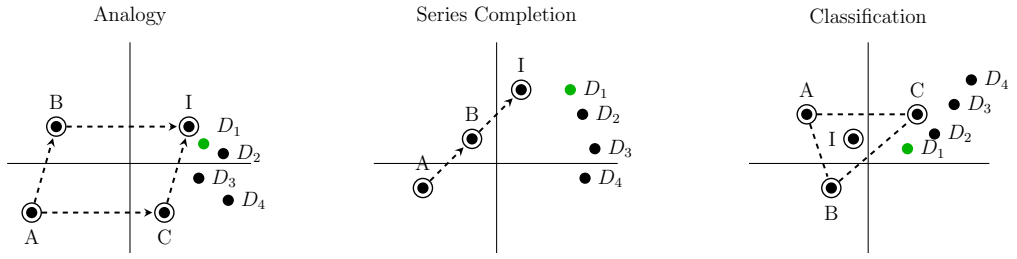

Figure 1: Inductive reasoning in semantic space proposed in Sternberg and Gardner (1983). A, B, C are given, I is the ideal point and D are the choices. The correct answer is shaded green.

Formally, we view word embedding methods as performing *metric recovery*. This perspective is significantly different from current approaches. Instead of aiming for representations that exhibit specific semantic properties or that perform well at a particular task, we seek methods that recover the underlying metric of the hypothesized semantic space. The clearer foundation afforded by this perspective enables us to analyze word embedding algorithms in a principled task-independent fashion. In particular, we ask whether word embedding algorithms are able to recover the metric under specific scenarios. To this end, we unify existing word embedding algorithms as statistically consistent metric recovery methods under the theoretical assumption that co-occurrences arise from (metric) random walks over semantic spaces. The new setting also suggests a simple and direct recovery algorithm which we evaluate and compare against other embedding methods.

The main contributions of this work can be summarized as follows:

- We ground word embeddings in *semantic spaces* via log co-occurrence counts. We show that PMI (pointwise mutual information) relates linearly to human similarity assessments, and that nearest-neighbor statistics (centrality and reciprocity) are consistent with an Euclidean space hypothesis (Sections 2 and 3).
- In contrast to prior work (Arora et al., 2015), we take *metric recovery* as the key object of study, unifying existing algorithms as consistent metric recovery methods based on co-occurrence counts from simple Markov random walks over graphs and manifolds. This strong link to manifold estimation opens a promising direction for extensions of word embedding methods (Sections and 4 and 5).
- We propose and evaluate a new principled direct metric recovery algorithm that performs comparably to the existing state of the art on both word embedding and manifold learning tasks, and show that `GloVe` (Pennington et al., 2014) is closely related to the second-order Taylor expansion of our objective.
- We construct and make available two new inductive reasoning datasets[1]—series completion and classification—to extend the evaluation of word representations beyond analogies, and demonstrate that these tasks can be solved with vector operations on word embeddings as well (Examples in Table 1).

2 Word vectors and semantic spaces

Most current word embedding algorithms build on the *distributional hypothesis* (Harris, 1954) where similar contexts imply similar meanings so as to tie co-occurrences of words to their underlying meanings. The relationship between semantics and co-occurrences has also been studied in psychometrics and cognitive science (Rumelhart and Abrahamson, 1973; Sternberg and Gardner, 1983), often by means of free word association tasks and *semantic spaces*. The semantic spaces, in particular, provide a natural conceptual framework for continuous representations of words as vector spaces where semantically related words are close to each other. For example, the observation that word embeddings can solve analogies was already shown by Rumelhart and Abrahamson (1973) using vector representations of words derived from surveys of pairwise word similarity judgments.

A fundamental question regarding vector space models of words is whether an Euclidean vector

[1] http://web.mit.edu/thashim/www/supplement_materials.zip

Task	Prompt	Answer
Analogy	king:man::queen:?	woman
Series	penny:nickel:dime:?	quarter
Classification	horse:zebra:{deer, dog, fish}	deer

Table 1: Examples of the three inductive reasoning tasks proposed by Sternberg and Gardner (1983).

space is a valid representation of semantic concepts. There is substantial empirical evidence in favor of this hypothesis. For example, Rumelhart and Abrahamson (1973) showed experimentally that analogical problem solving with fictitious words and human mistake rates were consistent with an Euclidean space. Sternberg and Gardner (1983) provided further evidence supporting this hypothesis, proposing that general inductive reasoning was based upon operations in metric embeddings. Using the analogy, series completion and classification tasks shown in Table 1 as testbeds, they proposed that subjects solve these problems by finding the word closest (in semantic space) to an ideal point: the vertex of a parallelogram for analogies, a displacement from the last word in series completion, and the centroid in the case of classification (Figure 1).

We use semantic spaces as the prototypical structures that word embedding methods attempt to uncover, and we investigate the suitability of word co-occurrence counts for doing so. In the next section, we show that co-occurrences from large corpora indeed relate to semantic similarity assessments, and that the resulting metric is consistent with an Euclidean semantic space hypothesis.

3 The semantic space of log co-occurrences

Most word embedding algorithms are based on word co-occurrence counts. In order for such methods to uncover an underlying Euclidean semantic space, we must demonstrate that co-occurrences themselves are indeed consistent with some semantic space. We must relate co-occurrences to semantic similarity assessments, on one hand, and show that they can be embedded into a Euclidean metric space, on the other. We provide here empirical evidence for both of these premises.

We commence by demonstrating in Figure 2 that the pointwise mutual information (Church and Hanks, 1990) evaluated from co-occurrence

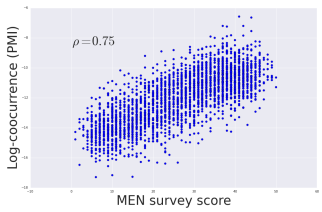

Figure 2: Normalized log co-occurrence (PMI) linearly correlates with human semantic similarity judgments (MEN survey).

counts has a strong linear relationship with semantic similarity judgments from survey data (Pearson's r=0.75).[2] However, this suggestive linear relationship does not by itself demonstrate that log co-occurrences (with normalization) can be used to define an Euclidean metric space.

Earlier psychometric studies have asked whether human semantic similarity evaluations are consistent with an Euclidean space. For example, Tversky and Hutchinson (1986) investigate whether concept representations are consistent with the *geometric sampling* (GS) model: a generative model in which points are drawn independently from a continuous distribution in an Euclidean space. They use two nearest neighbor statistics to test agreement with this model, and conclude that certain hierarchical vocabularies are not consistent with an Euclidean embedding. Similar results are observed by Griffiths et al. (2007). We extend this embeddability analysis to lexical co-occurrences and show that semantic similarity estimates derived from these are mostly consistent with an Euclidean space hypothesis.

The first test statistic for the GS model, the *centrality* C, is defined as

$$C = \frac{1}{n} \sum_{i=1}^{n} \Big(\sum_{j=1}^{n} N_{ij} \Big)^2$$

where $N_{ij} = 1$ iff i is j's nearest neighbor. Under the GS model (i.e. when the words are consistent with a Euclidean space representation), $C \leq 2$ with high probability as the number of words $n \to \infty$ regardless of the dimension or the underlying density (Tversky and Hutchinson, 1986). For metrically embeddable data, typical non-asymptotic values of C

[2] Normalizing the log co-occurrence with the unigram frequency taken to the $3/4$th power maximizes the linear correlation in Figure 2, explaining this choice of normalization in prior work (Levy and Goldberg, 2014a; Mikolov et al., 2013b).

Corpus	C	R_f
Free association	1.51	0.48
Wikipedia corpus	2.21	0.63
Word2vec corpus	2.24	0.73
GloVe corpus	2.66	0.62

Table 2: Semantic similarity data derived from multiple sources show evidence of embeddability

range between 1 and 3, while non-embeddable hierarchical structures have $C > 10$.

The second statistic, the *reciprocity fraction* R_f (Schwarz and Tversky, 1980; Tversky and Hutchinson, 1986), is defined as

$$R_f = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} N_{ij}N_{ji}$$

and measures the fraction of words that are their nearest neighbor's nearest neighbor. Under the GS model, this fraction should be greater than 0.5.[3]

Table 2 shows the two statistics computed on three popular large corpora and a free word association dataset (see Section 6 for details). The nearest neighbor calculations are based on PMI. The results show a surprisingly high agreement on all three statistics for all corpora, with C and R_f contained in small intervals: $C \in [2.21, 2.66]$ and $R_f \in [0.62, 0.73]$. These results are consistent with Euclidean semantic spaces and the GS model in particular. The largest violators of C and R_f are consistent with Tversky's analysis: the word with the largest centrality in the non-stopword Wikipedia corpus is 'the', whose inclusion would increase C to 3.46 compared to 2.21 without it. Tversky's original analysis of semantic similarities argued that certain words, such as superordinate and function words, could not be embedded. Despite such specific exceptions, we find that for an appropriately normalized corpus, the majority of words are consistent with the GS model, and therefore can be represented meaningfully as vectors in Euclidean space.

The results of this section are an important step towards justifying the use of word co-occurrence counts as the central object of interest for semantic vector representations of words. We have shown that they are *empirically* related to a human notion of semantic similarity and that they are metrically embeddable, a desirable condition if we expect word vectors derived from them to truly behave as elements of a metric space. This, however, does not yet fully justify their use to derive semantic representations. The missing piece is to formalize the connection between these co-occurrence counts and some intrinsic notion of semantics, such as the semantic spaces described in Section 2. In the next two sections, we establish this connection by framing word embedding algorithms that operate on co-occurrences as metric recovery methods.

[3]Both R and C are asymptotically dimension independent because they rely only on the single nearest neighbor. Estimating the latent dimensionality requires other measures and assumptions (Kleindessner and von Luxburg, 2015).

4 Semantic spaces and manifolds

We take a broader, unified view on metric recovery of semantic spaces since the notion of semantic spaces and the associated parallelogram rule for analogical reasoning extend naturally to objects other than words. For example, images can be approximately viewed as points in an Euclidean semantic space by representing them in terms of their underlying degrees of freedom (e.g. orientation, illumination). Thus, questions about the underlying semantic spaces and how they can be recovered should be related.

The problem of recovering an intrinsic Euclidean coordinate system over objects has been specifically addressed in manifold learning. For example, methods such as Isomap (Tenenbaum et al., 2000) reconstitute an Euclidean space over objects (when possible) based on local comparisons. Intuitively, these methods assume that naive distance metrics such as the L_2 distance over pixels in an image may be meaningful only when images are very similar. Longer distances between objects are evaluated through a series of local comparisons. These longer distances—*geodesic* distances over the manifold—can be approximated by shortest paths on a neighborhood graph. If we view the geodesic distances on the manifold (represented as a graph) as semantic distances, then the goal is to isometrically embed these distances into an Euclidean space. Tenenbaum (1998) showed that such isometric embeddings of image manifolds can be used to solve "visual analogies" via the parallelogram rule.

Typical approaches to manifold learning as dis-

cussed above differ from word embedding in terms of how the semantic distances between objects are extracted. Word embeddings approximate semantic distances between words using the negative log co-occurrence counts (Section 3), while manifold learning approximates semantic distances using neighborhood graphs built from local comparisons of the original, high-dimensional points. Both views seek to estimate a latent geodesic distance.

In order to study the problem of metric recovery from co-occurrence counts, and to formalize the connection between word embedding and manifold learning, we introduce a simple random walk model over the underlying objects (e.g. words or images). This toy model permits us to establish clean consistency results for recovery algorithms. We emphasize that while the random walk is introduced over the words, it is not intended as a model of language but rather as a tool to understand the recovery problem.

4.1 Random walk model

Consider now a simple metric random walk X_t over words where the probability of transitioning from word i to word j is given by

$$P(X_t = j | X_{t-1} = i) = h(\tfrac{1}{\sigma}||x_i - x_j||_2^2) \qquad (1)$$

Here $||x_i - x_j||_2^2$ is the Euclidean distance between words in the underlying semantic space to be recovered, and h is some unknown, sub-Gaussian function linking semantic similarity to co-occurrence.[4]

Under this model, the log frequency of occurrences of word j immediately after word i will be proportional to $\log(h(||x_i - x_j||_2^2/\sigma))$ as the corpus size grows large. Here we make the surprising observation that if we consider co-occurrences over a sufficiently large window, the log co-occurrence instead converges to $-||x_i - x_j||_2^2/\sigma$, i.e. it directly relates to the underlying metric. Intuitively, this result is an analog of the central limit theorem for random walks. Note that, for this reason, we do not need to know the link function h.

Formally, given an m-token corpus consisting of sentences generated according to Equation 1 from a vocabulary of size n, let $C_{ij}^{m,n}(t_n)$ be the number of times word j occurs t_n steps after word i in the corpus.[5] We can show that there exist unigram normalizers $a_i^{m,n}, b_i^{m,n}$ such that the following holds:

Lemma 1. *Given a corpus generated by Equation 1 there exists a_i and b_j such that simultaneously over all i, j:*

$$\lim_{m,n\to\infty} -\log(C_{ij}^{m,n}(t_n)) - a_i^{m,n} - b_j^{m,n} \to ||x_i - x_j||_2^2.$$

We defer the precise statement and conditions of Lemma 1 to Corollary 6. Conceptually, this limiting[6] result captures the intuition that while one-step transitions in a sentence may be complex and include non-metric structure expressed in h, co-occurrences over large windows relate directly to the latent semantic metric. For ease of notation, we henceforth omit the corpus and vocabulary size descriptors m, n (using C_{ij}, a_i, and b_j in place of $C_{ij}^{m,n}(t_n)$, $a_i^{m,n}$, and $b_j^{m,n}$), since in practice the corpus is large but fixed.

Lemma 1 serves as the basis for establishing consistency of recovery for word embedding algorithms (next section). It also allows us to establish a precise link between between manifold learning and word embedding, which we describe in the remainder of this section.

4.2 Connection to manifold learning

Let $\{v_1 \dots v_n\} \in \mathbb{R}^D$ be points drawn i.i.d. from a density p, where D is the dimension of observed inputs (e.g. number of pixels, in the case of images), and suppose that these points lie on a manifold $\mathcal{M} \subset \mathbb{R}^D$ that is isometrically embeddable into $d < D$ dimensions, where d is the intrinsic dimensionality of the data (e.g. coordinates representing illumination or camera angle in the case of images). The problem of manifold learning consists of finding an embedding of $v_1 \dots v_n$ into $\mathbb{R}^d$ that preserves the structure of $\mathcal{M}$ by approximately preserving the distances between points along this manifold. In light

[4] This toy model ignores the role of syntax and function words, but these factors can be included as long as the moment bounds originally derived in Hashimoto et al. (2015b) remain fulfilled.

[5] The window-size t_n depends on the vocabulary size to ensure that all word pairs have nonzero co-occurrence counts in the limit of large vocabulary and corpus. For details see the definition of g_n in Appendix A.

[6] In Lemma 1, we take $m \to \infty$ (growing corpus size) to ensure all word pairs appear sufficiently often, and $n \to \infty$ (growing vocabulary) to ensure that every point in the semantic space has a nearby word.

of Lemma 1, this problem can be solved with word embedding algorithms in the following way:

1. Construct a neighborhood graph (e.g. connecting points within a distance ε) over $\{v_1 \dots v_n\}$.
2. Record the vertex sequence of a simple random walk over these graphs as a sentence, and concatenate these sequences initialized at different nodes into a corpus.
3. Use a word embedding method on this corpus to generate d-dimensional vector representations of the data.

Under the conditions of Lemma 1, the negative log co-occurrences over the vertices of the neighborhood graph will converge, as $n \to \infty$, to the geodesic distance (squared) over the manifold. In this case we will show that the globally optimal solutions of word embedding algorithms recover the low dimensional embedding (Section 5).[7]

5 Recovering semantic distances with word embeddings

We now show that, under the conditions of Lemma 1, three popular word embedding methods can be viewed as doing metric recovery from co-occurrence counts. We use this observation to derive a new, simple word embedding method inspired by Lemma 1.

5.1 Word embeddings as metric recovery

GloVe The Global Vectors (`GloVe`) (Pennington et al., 2014) method for word embedding optimizes the objective function

$$\min_{\widehat{x},\widehat{c},a,b} \sum_{i,j} f(C_{ij})(2\langle \widehat{x}_i, \widehat{c}_j \rangle + a_i + b_j - \log(C_{ij}))^2$$

with $f(C_{ij}) = \min(C_{ij}, 100)^{3/4}$. If we rewrite the bias terms as $a_i = \widehat{a}_i - ||\widehat{x}_i||_2^2$ and $b_j = \widehat{b}_j - ||\widehat{c}_j||_2^2$, we obtain the equivalent representation:

$$\min_{\widehat{x},\widehat{c},\widehat{a},\widehat{b}} \sum_{i,j} f(C_{ij})(-\log(C_{ij}) - ||\widehat{x}_i - \widehat{c}_j||_2^2 + \widehat{a}_i + \widehat{b}_j))^2.$$

Together with Lemma 1, we recognize this as a weighted multidimensional scaling (MDS) objective with weights $f(C_{ij})$. Splitting the word vector $\widehat{x}_i$ and context vector $\widehat{c}_i$ is helpful in practice but not necessary under the assumptions of Lemma 1 since the true embedding $\widehat{x}_i = \widehat{c}_i = x_i/\sigma$ and $\widehat{a}_i, \widehat{b}_i = 0$ is a global minimum whenever $\dim(\widehat{x}) = d$. In other words, `GloVe` can recover the true metric provided that we set d correctly.

word2vec The skip-gram model of `word2vec` approximates a softmax objective:

$$\min_{\widehat{x},\widehat{c}} \sum_{i,j} C_{ij} \log\left(\frac{\exp(\langle \widehat{x}_i, \widehat{c}_j \rangle)}{\sum_{k=1}^n \exp(\langle \widehat{x}_i, \widehat{c}_k \rangle)}\right).$$

Without loss of generality, we can rewrite the above with a bias term b_j by making $\dim(\widehat{x}) = d + 1$ and setting one of the dimensions of $\widehat{x}$ to 1. By redefining the bias $\widehat{b}_j = b_j - ||\widehat{c}_j||_2^2/2$, we see that `word2vec` solves

$$\min_{\widehat{x},\widehat{c},\widehat{b}} \sum_{i,j} C_{ij} \log\left(\frac{\exp(-\frac{1}{2}||\widehat{x}_i - \widehat{c}_j||_2^2 + \widehat{b}_j)}{\sum_{k=1}^n \exp(-\frac{1}{2}||\widehat{x}_i - \widehat{c}_k||_2^2 + \widehat{b}_k)}\right).$$

Since according to Lemma 1 $C_{ij}/\sum_{k=1}^n C_{ik}$ approaches $\frac{\exp(-|||x_i - x_j||_2^2/\sigma^2)}{\sum_{k=1}^n \exp(-|||x_i - x_k||_2^2/\sigma^2)}$, this is the stochastic neighbor embedding (SNE) (Hinton and Roweis, 2002) objective weighted by $\sum_{k=1}^n C_{ik}$. The global optimum is achieved by $\widehat{x}_i = \widehat{c}_i = x_i(\sqrt{2}/\sigma)$ and $\widehat{b}_j = 0$ (see Theorem 8). The negative sampling approximation used in practice behaves much like the SVD approach of Levy and Goldberg (2014b), and by applying the same stationary point analysis as they do, we show that in the absence of a bias term the true embedding is a global minimum under the additional assumption that $||x_i||_2^2(2/\sigma^2) = \log(\sum_j C_{ij}/\sqrt{\sum_{ij} C_{ij}})$ (Hinton and Roweis, 2002).

SVD The method of Levy and Goldberg (2014b) uses the log PMI matrix defined in terms of the unigram frequency C_i as:

$$M_{ij} = \log(C_{ij}) - \log(C_i) - \log(C_j) + \log\Big(\sum_j C_j\Big)$$

and computes the SVD of the shifted and truncated matrix: $(M_{ij} + \tau)_+$ where τ is a truncation parameter to keep M_{ij} finite. Under the limit of Lemma 1, the corpus is sufficiently large that no truncation is necessary (i.e. $\tau = -\min(M_{ij}) < \infty$). We will

[7] This approach of applying random walks and word embeddings to general graphs has already been shown to be surprisingly effective for social networks (Perozzi et al., 2014), and demonstrates that word embeddings serve as a general way to connect metric random walks to embeddings.

recover the underlying embedding if we additionally assume $\frac{1}{\sigma}||x_i||_2^2 = \log(C_i/\sqrt{\sum_j C_j})$ via the law of large numbers since $M_{ij} \to \langle x_i, x_j \rangle$ (see Theorem 7). Centering the matrix M_{ij} before obtaining the SVD would relax the norm assumption, resulting exactly in classical MDS (Sibson, 1979).

5.2 Metric regression from log co-occurrences

We have shown that by through simple reparameterizations and use of Lemma 1, existing embedding algorithms can be interpreted as consistent metric recovery methods. However, the same Lemma suggests a more direct regression method for recovering the latent coordinates, which we propose here. This new embedding algorithm serves as a litmus test for our metric recovery paradigm.

Lemma 1 describes a log-linear relationship between distance and co-occurrences. The canonical way to fit this relationship would be to use a generalized linear model, where the co-occurrences follow a negative binomial distribution $C_{ij} \sim \text{NegBin}(\theta, p)$, where $p = \theta/[\theta + \exp(-\frac{1}{2}||x_i - x_j||_2^2 + a_i + b_j)]$. Under this overdispersed log linear model,

$$\mathbb{E}[C_{ij}] = \exp(-\tfrac{1}{2}||x_i - x_j||_2^2 + a_i + b_j)$$
$$\text{Var}(C_{ij}) = \mathbb{E}[C_{ij}]^2/\theta + \mathbb{E}[C_{ij}]$$

Here, the parameter θ controls the contribution of large C_{ij}, and is akin to GloVe's $f(C_{ij})$ weight function. Fitting this model is straightforward if we define the log-likelihood in terms of the expected rate $\lambda_{ij} = \exp(-\frac{1}{2}||x_i - x_j||_2^2 + a_i + b_j)$ as:

$$\text{LL}(x, a, b, \theta) = \sum_{i,j} \theta \log(\theta) - \theta \log(\lambda_{ij} + \theta) +$$
$$C_{ij} \log\left(1 - \tfrac{\theta}{\lambda_{ij}+\theta}\right) + \log\left(\tfrac{\Gamma(C_{ij}+\theta)}{\Gamma(\theta)\Gamma(C_{ij}+1)}\right)$$

To generate word embeddings, we minimize the negative log-likelihood using stochastic gradient descent. The implementation mirrors that of GloVe and randomly selects word pairs i, j and attracts or repulses the vectors $\widehat{x}$ and $\widehat{c}$ in order to achieve the relationship in Lemma 1. Implementation details are provided in Appendix C.

Relationship to GloVe The overdispersion parameter θ in our metric regression model sheds light on the role of GloVe's weight function $f(C_{ij})$. Taking the Taylor expansion of the log-likelihood at $\log(\lambda_{ij}) \approx -\log(C_{ij})$, we have

$$\text{LL}(x, a, b, \theta) = \sum_{ij} k_{ij} - \tfrac{C_{ij}\theta}{2(C_{ij}+\theta)}(u_{ij})^2 + o\big((u_{ij})^3\big),$$

where $u_{ij} = (\log \lambda_{ij} - \log C_{ij})$ and k_{ij} does not depend on x. Note the similarity of the second order term with the GloVe objective. As C_{ij} grows, the weight functions $\frac{C_{ij}\theta}{2(C_{ij}+\theta)}$ and $f(C_{ij}) = \max(C_{ij}, x_{max})^{3/4}$ converge to $\theta/2$ and x_{max} respectively, down-weighting large co-occurrences.

6 Empirical validation

We will now experimentally validate two aspects of our theory: the semantic space hypothesis (Sections 2 and 3), and the correspondence between word embedding and manifold learning (Sections 4 and 5). Our goal with this empirical validation is not to find the absolute best method and evaluation metric for word embeddings, which has been studied before (e.g. Levy et al. (2015)). Instead, we provide empirical evidence in favor of the semantic space hypothesis, and show that our simple algorithm for metric recovery is competitive with state-of-the-art on both semantic induction tasks and manifold learning. Since metric regression naturally operates over integer co-occurrences, we use co-occurrences over unweighted windows for this and—for fairness—for the other methods (see Appendix C for details).

6.1 Datasets

Corpus and training: We used three different corpora for training: a Wikipedia snapshot of 03/2015 (2.4B tokens), the original word2vec corpus (Mikolov et al., 2013a) (6.4B tokens), and a combination of Wikipedia with Gigaword5 emulating GloVe's corpus (Pennington et al., 2014) (5.8B tokens). We preprocessed all corpora by removing punctuation, numbers and lower-casing all the text. The vocabulary was restricted to the 100K most frequent words in each corpus. We trained embeddings using four methods: word2vec, GloVe, randomized SVD,[8] and metric regression (referred to as *regression*). Full implementation details are provided in the Appendix.

[8] We used randomized, rather than full SVD due to the difficulty of scaling SVD to this problem size. For performance of full SVD factorizations see Levy et al. (2015).

	Google Semantic		Google Syntactic		Google Total		SAT		Classification		Sequence	
Method	L_2	Cos	L_2	Cos	L_2	Cos	L_2	Cos	L_2	Cos	L_2	Cos
Regression	**75.5**	**78.4**	**70.9**	70.8	**72.6**	**73.7**	39.2	37.8	**87.6**	**84.6**	**58.3**	**59.0**
GloVe	71.1	76.4	68.6	71.9	69.6	**73.7**	36.9	35.5	74.6	80.1	53.0	58.9
SVD	50.9	58.1	51.4	52.0	51.2	54.3	32.7	24.0	71.6	74.1	49.4	47.6
word2vec	71.4	73.4	**70.9**	**73.3**	71.1	73.3	**42.0**	**42.0**	76.4	**84.6**	54.4	56.2

Table 3: Accuracies on Google, SAT analogies and on two new inductive reasoning tasks.

	Manifold Learning	Word Embedding
Semantic	**83.3**	70.7
Syntactic	8.2	**76.9**
Total	51.4	**73.4**

Table 4: Semantic similarity alone can solve the Google analogy tasks

For fairness we fix all hyperparameters, and develop and test the code for metric regression exclusively on the first 1GB subset of the wiki dataset. For open-vocabulary tasks, we restrict the set of answers to the top 30K words, since this improves performance while covering the majority of the questions. In the following, we show performance for the `GloVe` corpus throughout but include results for all corpora along with our code package.

Evaluation tasks: We test the quality of the word embeddings on three types of inductive tasks: analogies, sequence completion and classification (Figure 1). For the analogies, we used the standard open-vocabulary analogy task of Mikolov et al. (2013a) (henceforth denoted Google), consisting of 19,544 semantic and syntactic questions. In addition, we use the more difficult SAT analogy dataset (version 3) (Turney and Littman, 2005), which contains 374 questions from actual exams and guidebooks. Each question consists of 5 exemplar pairs of words *word1:word2*, where all the pairs hold the same relation. The task is to pick from among another five pairs of words the one that best fits the category implicitly defined by the exemplars.

Inspired by Sternberg and Gardner (1983), we propose two new difficult inductive reasoning tasks beyond analogies to verify the semantic space hypothesis: sequence completion and classification. As described in Section 2, the former involves choosing the next step in a semantically coherent sequence of words (e.g. $hour, minute, \ldots$), and the latter consists of selecting an element within the same category out of five possible choices. Given the lack of publicly available datasets, we generated our own questions using WordNet (Fellbaum, 1998) relations and word-word PMI values. These datasets were constructed before training the embeddings, so as to avoid biasing them towards any one method.

For the classification task, we created in-category words by selecting words from WordNet relations associated to root words, from which we pruned to four words based on PMI-similarity to the other words in the class. Additional options for the multiple choice questions were created searching over words related to the root by a different relation type, and selecting those most similar to the root.

For the sequence completion task, we obtained WordNet trees of various relation types, and pruned these based on similarity to the root word to obtain the sequence. For the multiple-choice questions, we proceeded as before to selected additional (incorrect) options of a different relation type to the root.

After pruning, we obtain 215 classification questions and 220 sequence completion questions, of which 51 are open-vocabulary and 169 are multiple choice. These two new datasets will be released.

6.2 Results on inductive reasoning tasks

Solving analogies using survey data alone: We demonstrate that, surprisingly, word embeddings trained directly on semantic similarity derived from survey data can solve analogy tasks. Extending a study by Rumelhart and Abrahamson (1973), we use a free-association dataset (Nelson et al., 2004) to construct a similarity graph, where vertices correspond to words and the weights w_{ij} are given by the number of times word j was considered most similar to word i in the survey. We take the largest connected component of this graph (consisting of 4845 words and 61570 weights) and embed it using Isomap for which squared edge distances are defined as $-\log(w_{ij}/\max_{kl}(w_{kl}))$. We use the resulting vectors as word embeddings to solve the Google

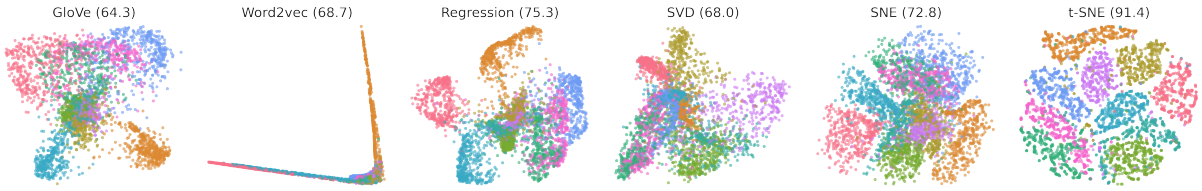

Figure 3: Dimensionality reduction using word embedding and manifold learning. Performance is quantified by percentage of 5-nearest neighbors sharing the same digit label.

analogy task. The results in Table 4 show that embeddings obtained with Isomap on survey data can outperform the corpus based metric regression vectors on semantic, but not syntactic tasks. We hypothesize this is due to the fact that free-association surveys capture semantic, but not syntactic similarity between words.

Analogies: The results on the Google analogies shown in Table 3 demonstrate that our proposed framework of metric regression and L_2 distance is competitive with the baseline of `word2vec` with cosine distance. The performance gap across methods is small and fluctuates across corpora, but metric regression consistently outperforms `GloVe` on most tasks and outperforms all methods on semantic analogies, while `word2vec` does better on syntactic categories. For the SAT dataset, the L_2 distance performs better than the cosine similarity, and we find `word2vec` to perform best, followed by metric regression. The results on these two analogy datasets show that directly embedding the log co-occurrence metric and taking L_2 distances between vectors is competitive with current approaches to analogical reasoning.

Sequence and classification tasks: As predicted by the semantic field hypothesis, word embeddings perform well on the two novel inductive reasoning tasks (Table 3). Again, we observe that the metric recovery with metric regression coupled with L_2 distance consistently performs as well as and often better than the current state-of-the-art word embedding methods on these two additional semantic datasets.

6.3 Word embeddings can embed manifolds

In Section 4 we proposed a reduction for solving manifold learning problems with word embeddings which we show achieves comparable performance to manifold learning methods. We now test this relation by performing nonlinear dimensionality reduction on the MNIST digit dataset, reducing from $D = 256$ to two dimensions. Using a four-thousand image subset, we construct a k-nearest neighbor graph ($k = 20$) and generate 10 simple random walks of length 200 starting from each vertex in the graph, resulting in 40,000 sentences of length 200 each. We compare the four word embedding methods against standard dimensionality reduction methods: PCA, Isomap, SNE and, t-SNE. We evaluate the methods by clustering the resulting low-dimensional data and computing cluster purity, measured using the percentage of 5-nearest neighbors having the same digit label. The resulting embeddings, shown in Fig. 3, demonstrate that metric regression is highly effective at this task, outperforming metric SNE and beaten only by t-SNE (91% cluster purity), which is a visualization method specifically designed to preserve cluster separation. All word embedding methods including SVD (68%) embed the MNIST digits remarkably well and outperform baselines of PCA (48%) and Isomap (49%).

7 Discussion

Our work recasts word embedding as a metric recovery problem pertaining to the underlying semantic space. We use co-occurrence counts from random walks as a theoretical tool to demonstrate that existing word embedding algorithms are consistent metric recovery methods. Our direct regression method is competitive with the state of the art on various semantics tasks, including two new inductive reasoning problems of series completion and classification.

Our framework highlights the strong interplay and common foundation between word embedding methods and manifold learning, suggesting several avenues for recovering vector representations of phrases and sentences via properly defined Markov processes and their generalizations.

Appendix

A Metric recovery from Markov processes on graphs and manifolds

Consider an infinite sequence of points $\mathcal{X}_n = \{x_1, \ldots, x_n\}$, where x_i are sampled i.i.d. from a density $p(x)$ over a compact Riemannian manifold equipped with a geodesic metric ρ. For our purposes, $p(x)$ should have a bounded log-gradient and a strict lower bound p_0 over the manifold. The random walks we consider are over *unweighted spatial graphs* defined as

Definition 2 (Spatial graph)**.** *Let $\sigma_n : \mathcal{X}_n \to \mathbb{R}_{>0}$ be a local scale function and $h : \mathbb{R}_{\geq 0} \to [0,1]$ a piecewise continuous function with sub-Gaussian tails. A spatial graph G_n corresponding to σ_n and h is a random graph with vertex set $\mathcal{X}_n$ and a directed edge from x_i to x_j with probability $p_{ij} = h(\rho(x_i, x_j)^2/\sigma_n(x_i)^2)$.*

Simple examples of spatial graphs where the connectivity is not random include the ε ball graph ($\sigma_n(x) = \varepsilon$) and the k-nearest neighbor graph ($\sigma_n(x)$ =distance to k-th neighbor).

Log co-occurrences and the geodesic will be connected in two steps. (1) we use known results to show that a simple random walk over the spatial graph, properly scaled, behaves similarly to a diffusion process; (2) the log-transition probability of a diffusion process will be related to the geodesic metric on a manifold.

(1) The limiting random walk on a graph: Just as the simple random walk over the integers converges to a Brownian motion, we may expect that under specific constraints the simple random walk X_t^n over the graph G_n will converge to some well-defined continuous process. We require that the scale functions converge to a continuous function $\bar{\sigma}$ ($\sigma_n(x) g_n^{-1} \xrightarrow{a.s.} \bar{\sigma}(x)$); the size of a single step vanish ($g_n \to 0$) but contain at least a polynomial number of points within $\sigma_n(x)$ ($g_n n^{\frac{1}{d+2}} \log(n)^{-\frac{1}{d+2}} \to \infty$). Under this limit, our assumptions about the density $p(x)$, and regularity of the transitions[9], the following holds:

Theorem 3 ((Hashimoto et al., 2015b; Ting et al., 2011))**.** *The simple random walk X_t^n on G_n converges in Skorokhod space $\mathsf{D}([0,\infty), \overline{D})$ after a time scaling $\widehat{t} = t g_n^2$ to the Itô process $Y_{\widehat{t}}$ valued in $\mathsf{C}([0,\infty), \overline{D})$ as $X^n_{\widehat{t} g_n^{-2}} \to Y_{\widehat{t}}$. The process $Y_{\widehat{t}}$ is defined over the normal coordinates of the manifold (D, g) with reflecting boundary conditions on D as*

$$dY_{\widehat{t}} = \nabla \log(p(Y_{\widehat{t}}))\overline{\sigma}(Y_{\widehat{t}})^2 d\widehat{t} + \overline{\sigma}(Y_{\widehat{t}}) dW_{\widehat{t}} \quad (2)$$

The equicontinuity constraint on the marginal densities of the random walk implies that the transition density for the random walk converges to its continuum limit.

Lemma 4 (Convergence of marginal densities)**.** *(Hashimoto et al., 2015a) Let x_0 be some point in our domain $\mathcal{X}_n$ and define the marginal densities $\widehat{q}_t(x) = \mathbb{P}(Y_t = x | Y_0 = x_0)$ and $q_{t_n}(x) = \mathbb{P}(X_t^n = x | X_0^n = x_0)$. If $t_n g_n^2 = \widehat{t} = \Theta(1)$, then under condition $(\star)$ and the results of Theorem 3 such that $X_t^n \to Y_t^n$ weakly, we have*

$$\lim_{n\to\infty} n q_{t_n}(x) = \widehat{q}_{\widehat{t}}(x) p(x)^{-1}.$$

(2) Log transition probability as a metric We may now use the stochastic process $Y_{\widehat{t}}$ to connect the log transition probability to the geodesic distance using Varadhan's large deviation formula.

Theorem 5 ((Varadhan, 1967; Molchanov, 1975))**.** *Let Y_t be a Itô process defined over a complete Riemann manifold (D, g) with geodesic distance $\rho(x_i, x_j)$ then*

$$\lim_{t\to 0} -t \log(\mathbb{P}(Y_t = x_j | Y_0 = x_i)) \to \rho(x_i, x_j)^2.$$

This estimate holds more generally for any space admitting a diffusive stochastic process (Saloff-Coste, 2010). Taken together, we finally obtain:

Corollary 6 (Varadhan's formula on graphs)**.** *For any δ,γ,n_0 there exists some $\widehat{t}$, $n > n_0$, and sequence b_j^n such that the following holds for the simple random walk X_t^n:*

$$\mathbb{P}\Big(\sup_{x_i, x_j \in \mathcal{X}_{n_0}} \Big| \widehat{t} \log(\mathbb{P}(X^n_{\widehat{t} g_n^{-2}} = x_j \mid X_0^n = x_i)) - \widehat{t} b_j^n - \rho_{\overline{\sigma}(x)}(x_i, x_j)^2 \Big| > \delta \Big) < \gamma$$

[9] For $t = \Theta(g_n^{-2})$, the marginal distribution $n\mathbb{P}(X_t | X_0)$ must be a.s. uniformly equicontinuous. For undirected spatial graphs, this is always true (Croydon and Hambly, 2008), but for directed graphs it is an open conjecture from (Hashimoto et al., 2015b).

Where $\rho_{\overline{\sigma}(x)}$ is the geodesic defined as

$$\rho_{\overline{\sigma}(x)}(x_i, x_j) = \min_{f \in C^1 : f(0)=x_i, f(1)=x_j} \int_0^1 \overline{\sigma}(f(t))dt$$

Proof. The proof is in two parts. First, by Varadhan's formula (Theorem 5, (Molchanov, 1975, Eq. 1.7)) for any $\delta_1 > 0$ there exists some $\widehat{t}$ such that:

$$\sup_{y,y' \in D} |-\widehat{t}\log(\mathbb{P}(Y_{\widehat{t}} = y'|Y_0 = y)) - \rho_{\overline{\sigma}(x)}(y', y)^2| < \delta_1$$

The uniform equicontinuity of the marginals implies their uniform convergence (Lemma S4), so for any $\delta_2 > 0$ and γ_0, there exists a n such that

$$\mathbb{P}(\sup_{x_j, x_i \in \mathcal{X}_{n_0}} |\mathbb{P}(Y_{\widehat{t}} = x_j|Y_0 = x_i) - np(x_j)\mathbb{P}(X^n_{g_n^{-2}\widehat{t}} = x_j|X^n_0 = x_i)| > \delta_2) < \gamma_0$$

By the lower bound on p and compactness of D, $\mathbb{P}(Y_{\widehat{t}}|Y_0)$ is lower bounded by some strictly positive constant c and we can apply uniform continuity of $\log(x)$ over (c, ∞) to get that for some δ_3 and γ,

$$\mathbb{P}\Big(\sup_{x_j, x_i \in \mathcal{X}_{n_0}} |\log(\mathbb{P}(Y_{\widehat{t}} = x_j|Y_0 = x_i)) - \log(np(x_j)) - \log(\mathbb{P}(X^n_{g_n^{-2}\widehat{t}} = x_j|X^n_0 = x_i))| > \delta_3\Big) < \gamma. \quad (3)$$

Finally we have the bound,

$$\mathbb{P}\Big(\sup_{x_i, x_j \in \mathcal{X}_{n_0}} |-\widehat{t}\log(\mathbb{P}(X^n_{g_n^{-2}\widehat{t}} = x_j|X^n_0 = x_i)) - \widehat{t}\log(np(x_j)) - \rho_{\overline{\sigma}(x)}(x_i, x_j)^2| > \delta_1 + \widehat{t}\delta_3\Big) < \gamma$$

To combine the bounds, given some δ and γ, set $b^n_j = \log(np(x_j))$, pick $\widehat{t}$ such that $\delta_1 < \delta/2$, then pick n such that the bound in Eq. 3 holds with probability γ and error $\delta_3 < \delta/(2\widehat{t})$. □

B Consistency proofs for word embedding

Lemma 7 (Consistency of SVD)**.** *Assume the norm of the latent embedding is proportional to the unigram frequency*

$$||x_i||/\sigma^2 = C_i/(\sum_j C_j)^{\frac{1}{2}}$$

Under these conditions, Let $\widehat{X}$ be the embedding derived from the SVD of M_{ij} as

$$2\widehat{X}\widehat{X}^T = M_{ij} = \log(C_{ij}) - \log\Big(C_i\Big) - \log\Big(C_j\Big) + \log\Big(\sum_i C_i\Big) + \tau.$$

Then there exists a τ such that this embedding is close to the true embedding under the same equivalence class as Lemma S7

$$\mathbb{P}\Big(\sum_i ||A\widehat{x}_i/\sigma^2 - x_j||_2^2 > \delta\Big) < \varepsilon.$$

Proof. By Corollary 6 for any $\delta_1 > 0$ and $\varepsilon_1 > 0$ there exists a m such that

$$P(\sup_{i,j} | - \log(C_{ij}) - (||x_i - x_j||_2^2/\sigma^2) - \log(mc)| > \delta_1) < \varepsilon_1.$$

Now additionally, if $C_i/\sqrt{\sum_j C_j} = ||x_i||^2/\sigma^2$ then we can rewrite the above bound as

$$P(\sup_{i,j} |\log(C_{ij}) - \log(C_i) - \log(C_j) + \log(\sum_i C_i) - 2\langle x_i, x_j\rangle/\sigma^2 - \log(mc)| > \delta_1) < \varepsilon_1.$$

and therefore,

$$P(\sup_{i,j} |M_{ij} - 2\langle x_i, x_j\rangle/\sigma^2 - \log(mc)| > \delta_1) < \varepsilon_1.$$

Given that the dot product matrix has error at most δ_1, the resulting embedding it known to have at most $\sqrt{\delta_1}$ error (Sibson, 1979).

This completes the proof, since we can pick $\tau = -\log(mc)$, $\delta_1 = \delta^2$ and $\varepsilon_1 = \varepsilon$. □

Theorem 8 (Consistency of softmax/word2vec)**.** *Define the softmax objective function with bias as*

$$g(\widehat{x}, \widehat{c}, \widehat{b}) = \sum_{ij} C_{ij} \log \frac{\exp(-||\widehat{x}_i - \widehat{c}_j||_2^2 + \widehat{b}_j)}{\sum_{k=1}^n \exp(-||\widehat{x}_i - \widehat{c}_k||_2^2 + \widehat{b}_k)}$$

Define $\overline{x}_m, \overline{c}_m, \overline{b}_m$ as the global minima of the above objective function for a co-occurrence C_{ij} over a corpus of size m. For any $\varepsilon > 0$ and $\delta > 0$ there exists some m such that

$$\mathbb{P}(|g(\tfrac{x}{\sigma}, \tfrac{x}{\sigma}, 0) - g(\overline{x}_m, \overline{c}_m, \overline{b}_m)| > \delta) < \varepsilon$$

Proof. By differentiation, any objective of the form

$$\min_{\lambda_{ij}} C_{ij} \log\left(\frac{\exp(-\lambda_{ij})}{\sum_k \exp(-\lambda_{ik})}\right)$$

has the minima $\lambda^*_{ij} = -\log(C_{ij}) + a_i$ up to un-identifiable a_i with objective function value

$C_{ij} \log(C_{ij}/\sum_k C_{ik})$. This gives a global function lower bound

$$g(\overline{x}_m, \overline{c}_m, \overline{b}_m) \geq \sum_{ij} C_{ij} \log\left(\frac{C_{ij}}{\sum_k C_{ik}}\right) \quad (4)$$

Now consider the function value of the true embedding $\frac{x}{\sigma}$;

$$g(\tfrac{x}{\sigma}, \tfrac{x}{\sigma}, 0) = \sum_{ij} C_{ij} \log \frac{\exp(-\frac{1}{\sigma^2}||x_i - x_j||_2^2)}{\sum_k \exp(-\frac{1}{\sigma^2}||x_i - x_k||_2^2)}$$

$$= \sum_{ij} C_{ij} \log\left(\frac{\exp(\log(C_{ij}) + \delta_{ij} + a_i)}{\sum_k \exp(\log(C_{ik}) + \delta_{ik} + a_i)}\right).$$

We can bound the error variables δ_{ij} using Corollary 6 as $\sup_{ij} |\delta_{ij}| < \delta_0$ with probability ε_0 for sufficiently large m with $a_i = \log(m_i) - \log(\sum_{k=1}^n \exp(-||x_i - x_k||_2^2/\sigma^2))$.

Taking the Taylor expansion at $\delta_{ij} = 0$, we have

$$g(\tfrac{x}{\sigma}, \tfrac{x}{\sigma}, 0) =$$

$$\sum_{ij} C_{ij} \log \tfrac{C_{ij}}{\sum_k C_{ik}} + \sum_{l=1}^{n} \tfrac{C_{il}}{\sum_k C_{ik}\delta_{il}} + o(||\delta||_2^2)$$

By the law of large numbers of C_{ij},

$$\mathbb{P}\Big(\Big|g(\tfrac{x}{\sigma}, \tfrac{x}{\sigma}, 0) - \sum_{ij} C_{ij} \log\left(\tfrac{C_{ij}}{\sum_k C_{ik}}\right)\Big| > n\delta_0\Big) < \varepsilon_0$$

which combined with (4) yields

$$\mathbb{P}(|g(\tfrac{x}{\sigma}, \tfrac{x}{\sigma}, 0) - g(\overline{x}, \overline{c}, \overline{b})| > n\delta_0) < \varepsilon_0.$$

To obtain the original theorem statement, take m to fulfil $\delta_0 = \delta/n$ and $\varepsilon_0 = \varepsilon$. □

Note that for `word2vec` with negative-sampling, applying the stationary point analysis of Levy and Goldberg (2014b) combined with the analysis in Lemma S7 shows that the true embedding is a global minimum.

C Empirical evaluation details

C.1 Implementation details

We used off-the-shelf implementations of `word2vec`[10] and `GloVe`[11]. The two other methods (randomized) SVD and regression embedding are both implemented on top of the `GloVe` codebase. We used 300-dimensional vectors and window size 5 in all models. Further details are provided below.

[10] http://code.google.com/p/word2vec

[11] http://nlp.stanford.edu/projects/glove

`word2vec`. We used the skip-gram version with 5 negative samples, 10 iterations, $\alpha = 0.025$ and frequent word sub-sampling with a parameter of 10^{-3}.

`GloVe`. We disabled `GloVe`'s corpus weighting, since this generally produced superior results. The default step-sizes results in NaN-valued embeddings, so we reduced them. We used $X_{\text{MAX}} = 100$, $\eta = 0.01$ and 10 iterations.

SVD. For the SVD algorithm of Levy and Goldberg (2014b), we use the `GloVe` co-occurrence counter combined with a parallel randomized projection SVD factorizer, based upon the `redsvd` library due to memory and runtime constraints.[12] Following Levy et al. (2015), we used the square root factorization, no negative shifts ($\tau = 0$ in our notation), and 50,000 random projections.

Regression Embedding. We use standard SGD with two differences. First, we drop co-occurrence values with probability proportional to $1 - C_{ij}/10$ when $C_{ij} < 10$, and scale the gradient, which resulted in training time speedups with no loss in accuracy. Second, we use an initial line search step combined with a linear step size decay by epoch. We use $\theta = 50$ and η is line-searched starting at $\eta = 10$.

C.2 Solving inductive reasoning tasks

The ideal point for a task is defined below:

- **Analogies**: Given A:B::C, the ideal point is given by $B - A + C$ (parallelogram rule).
- **Analogies (SAT)**: Given prototype A:B and candidates $C_1 : D_1 \ldots C_n : D_n$, we compare $D_i - C_i$ to the ideal point $B - A$.
- **Categories**: Given a category implied by $w_1, \ldots, w_n$, the ideal point is $I = \frac{1}{n}\sum_{i=1}^n w_i$.
- **Sequence**: Given sequence $w_1 : \cdots : w_n$ we compute the ideal as $I = w_n + \frac{1}{n}(w_n - w_1)$.

Once we have the ideal point I, we pick the answer as the word closest to I among the options, using L_2 or cosine distance. For the latter, we normalize I to unit norm before taking the cosine distance. For L_2 we do not apply any normalization.

[12] https://github.com/ntessore/redsvd-h

References

[Arora et al.2015] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2015. Random walks on context spaces: Towards an explanation of the mysteries of semantic word embeddings. *arXiv preprint arXiv:1502.03520*.

[Church and Hanks1990] Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29.

[Croydon and Hambly2008] David A Croydon and Ben M Hambly. 2008. Local limit theorems for sequences of simple random walks on graphs. *Potential Analysis*, 29(4):351–389.

[Fellbaum1998] Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*.

[Griffiths et al.2007] Thomas L Griffiths, Mark Steyvers, and Joshua B Tenenbaum. 2007. Topics in semantic representation. *Psychological review*, 114(2):211.

[Harris1954] Zellig S Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

[Hashimoto et al.2015a] Tatsunori Hashimoto, Yi Sun, and Tommi Jaakkola. 2015a. From random walks to distances on unweighted graphs. In *Advances in neural information processing systems*.

[Hashimoto et al.2015b] Tatsunori Hashimoto, Yi Sun, and Tommi Jaakkola. 2015b. Metric recovery from directed unweighted graphs. In *Artificial Intelligence and Statistics*, pages 342–350.

[Hinton and Roweis2002] Geoffrey E Hinton and Sam T Roweis. 2002. Stochastic neighbor embedding. In *Advances in neural information processing systems*, pages 833–840.

[Kleindessner and von Luxburg2015] Matthäus Kleindessner and Ulrike von Luxburg. 2015. Dimensionality estimation without distances. In *AISTATS*.

[Levy and Goldberg2014a] Omer Levy and Yoav Goldberg. 2014a. Linguistic Regularities in Sparse and Explicit Word Representations. *Proc. 18th Conf. Comput. Nat. Lang. Learn.*

[Levy and Goldberg2014b] Omer Levy and Yoav Goldberg. 2014b. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, pages 2177–2185.

[Levy et al.2015] Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

[Mikolov et al.2013a] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[Mikolov et al.2013b] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

[Molchanov1975] SA Molchanov. 1975. Diffusion processes and riemannian geometry. *Russian Mathematical Surveys*, 30(1):1.

[Nelson et al.2004] Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

[Pennington et al.2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12.

[Perozzi et al.2014] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM.

[Rumelhart and Abrahamson1973] David E Rumelhart and Adele A Abrahamson. 1973. A model for analogical reasoning. *Cognitive Psychology*, 5(1):1–28.

[Saloff-Coste2010] Laurent Saloff-Coste. 2010. The heat kernel and its estimates. *Probabilistic approach to geometry*, 57:405–436.

[Schwarz and Tversky1980] Gideon Schwarz and Amos Tversky. 1980. On the reciprocity of proximity relations. *Journal of Mathematical Psychology*, 22(3):157–175.

[Sibson1979] Robin Sibson. 1979. Studies in the robustness of multidimensional scaling: Perturbational analysis of classical scaling. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 217–229.

[Socher et al.2013] Richard Socher, John Bauer, Christopher D Manning, and Andrew Y Ng. 2013. Parsing with compositional vector grammars. In *In Proceedings of the ACL conference*. Citeseer.

[Sternberg and Gardner1983] Robert J Sternberg and Michael K Gardner. 1983. Unities in inductive reasoning. *Journal of Experimental Psychology: General*, 112(1):80.

[Tenenbaum et al.2000] Joshua B Tenenbaum, Vin De Silva, and John C Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

[Tenenbaum1998] Joshua B Tenenbaum. 1998. Mapping a manifold of perceptual observations. *Advances*

in neural information processing systems, pages 682–688.

[Ting et al.2011] Daniel Ting, Ling Huang, and Michael Jordan. 2011. An analysis of the convergence of graph laplacians. *arXiv preprint arXiv:1101.5435*.

[Turian et al.2010] Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. ACL.

[Turney and Littman2005] Peter D Turney and Michael L Littman. 2005. Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60(1-3):251–278.

[Tversky and Hutchinson1986] Amos Tversky and J Hutchinson. 1986. Nearest neighbor analysis of psychological spaces. *Psychological Review*, 93(1):3.

[Varadhan1967] Srinivasa RS Varadhan. 1967. Diffusion processes in a small time interval. *Communications on Pure and Applied Mathematics*, 20(4):659–685.