Word, graph and manifold embedding from Markov processes

Tatsunori B. Hashimoto
MIT CSAIL
thashim@csail.mit.edu

David Alvarez-Melis
MIT CSAIL
davidam@csail.mit.edu

Tommi S. Jaakkola
MIT CSAIL
tommi@mit.edu

This is an extended abstract of a preprint [6] containing full technical details.

1 Introduction

Continuous space models of words, objects, and signals have proven to be powerful tools for learning expressive representations of data, and been adopted across areas, from natural language processing to computer vision. Recently, there has been particular interest in word embeddings, largely due to their intriguing semantic properties [12] and their successful use as features for downstream NLP tasks [22, 19]. Embedding methods based on neural networks [3, 14, 12] have been at the forefront of this trend thanks to their simplicity, scalability and semantically-rich embeddings, but other nonparametric embedding methods have proven to share similar properties [9].

Despite their empirical success, understanding of word embeddings has lagged behind. Recent work has started to fill this gap, seeking better understanding of these representations, their properties, and associated algorithms [9, 4, 10, 1]. Yet, some questions are not fully answered yet. For example, it has been widely demonstrated that word embeddings can be used to solve analogy tasks. What remains to be explained is why: how is it that analytical reasoning, a complex cognitive process, can be replicated with simple operations on vector representations of words? We attempt to provide an explanation for this, by drawing a connection between the cognitive perspective of analogical reasoning, semantic similarities and the embedding of cooccurrence counts.

In this work we extend both the conceptual and theoretical understanding of word embeddings. First, we motivate them by examining the psychometric and cognitive basis for embeddings. In particular, we ground embeddings in *semantic spaces*, and revisit word vector representations derived from word association tasks, which were shown to be have similar linear structure as those shown by modern methods. Second, we propose a new theoretical framework for understanding word embeddings as a type of manifold learning. In contrast to prior work [1], 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, and propose a new algorithm which directly recovers an underlying metric.

2 Word vectors and semantic spaces

The conceptual motivation of most current word embedding relies largely on the *distributional hypothesis* [5]: words appearing in similar contexts tend to have similar meanings. Hence, word co-occurrence counts lie at the heart of all these approaches. But besides linguistics, word co-occurrences and their relationship with semantics also play a central role in psychometrics and cognitive science, where semantic similarity has long been studied [18, 20], often by means of free word association tasks and *semantic spaces*, i.e., vector spaces where semantically related words are close to each other. Yet, this rich literature seems to have been largely overlooked by recent work on word embeddings.

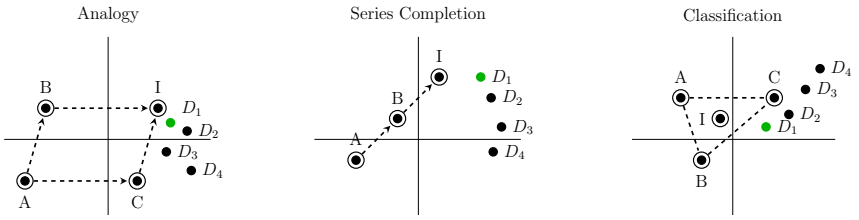

Figure 1: Sternberg's model for inductive reasoning in semantic space. A, B, C are given, I is the ideal point and D are the choices. The correct answer is shaded green. Adapted from [20].

Semantic spaces such as those used in the psychometrics literature provide a natural conceptual framework for continuous word representations. For one, the intriguing observation that word embeddings can be used to solve analogies has a natural explanation in this framework. In fact, this was already shown by Rumelhart and Abrahamson [18] using continuous word representations derived from semantic similarity surveys. The explanation provided there is that solving analogies amounts to a similarity judgment between the relations among two pairs of words. If these words are represented in a multidimensional euclidean space, then the most natural way of assessing this similarity is to compare the vectors between the two pairs of words. The question is thus whether a metric space is a valid representation of semantic concepts. There is significant empirical evidence supporting this. For example, it was shown in [18] that synthetic terms assigned to points in semantic space were used by subjects for solving analogies in the same way they used real words, and that human mistake rates followed an exponential decay in embedded distance from the true solution. Sternberg and Gardner provided further evidence supporting this hypothesis for analogical reasoning, proposing that general inductive reasoning was based upon operations in *metric embeddings* [20]. Using analogy, series completion and classification[1] tasks 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).

In this work, we use these cognitive semantic spaces as motivation for the underlying spaces that word embedding methods attempt to recover. Besides providing grounding from a cognitive perspective and offering an explanation for some of the properties of corpus based word embeddings, the link with the psychometric literature provides yet another advantage. It reminds us that there are other types of inductive reasoning besides analogical, which has recently dominated the evaluation of word embeddings. Tasks such as the series completion and classification [20] require similar operations on semantic entities, and thus a more robust evaluation scheme should also include those. Based on this observation, we propose two new inductive reasoning tasks, and demonstrate that word embeddings can be used to solve those too. For example, in the series completion task, given "body, arm, hand" we find the answer predicted by vector operations on word embeddings to be "fingers". We make these new datasets available to be used as benchmarks in addition to current popular analogy tasks.

3 Recovering semantic distances with word embedding

We illustrate the metric recovery properties of word embedding methods using a simple model proposed in the literature [2] and generalize the model in the next section. Our corpus consists of m total words across s sentences over a n word vocabulary where each word is given a coordinate in a latent word vector space $\{x_1, \ldots, x_n\} \in \mathbb{R}^d$. For each sentence s we consider a Markov random walk, $X_1, \ldots, X_{m_s}$, with the following transition function

$$\mathbb{P}(X_t = x_j | X_{t-1} = x_i) = \frac{\exp(-||x_i - x_j||_2^2/\sigma^2)}{\sum_{k=1}^{n} \exp(-||x_i - x_k||_2^2/\sigma^2)}. \quad (1)$$

[1] Choosing the word that fest fits a semantic category defined by a set of words.

Suppose we observe the Gaussian random walk (Eq. 1) over a corpus with m total words and define C_{ij} as the number of times for which $X_t = x_j$ and $X_{t-1} = x_i$.[2] By the Markov chain law of large numbers, as $m \to \infty$ where we take the limit as the number of sentences grows, $-\log\left(C_{ij}/\sum_{k=1}^n C_{ik}\right) \xrightarrow{p} ||x_i - x_j||_2^2/\sigma^2 + \log(Z_i)$ where $Z_i = \sum_{k=1}^n \exp(-|||x_i - x_k||_2^2/\sigma^2)$.

We first show that for the case of Eq 1, word embeddings recover the true latent embeddings x_i.

GloVe: The Global Vectors (`GloVe`) [16] 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 $f(C_{ij}) = \min(C_{ij}, 10)^{3/4}$. 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:

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

This is a weighted multidimensional scaling objective with weights $f(C_{ij})$. Splitting the word vector $\widehat{x}_i$ and context vector $\widehat{c}_i$ is helpful in practice to optimize this objective, but not necessary for our model since the true embedding $\widehat{x}_i = \widehat{c}_i = x_i/\sigma$ and $\widehat{a}_i, \widehat{b}_i = 0$ is one of the global minima whenever $\dim(\widehat{x}) = d$.

word2vec: The embedding algorithm `word2vec` approximates a softmax objective $\min_{\widehat{x},\widehat{c}} \sum_{i,j} C_{ij} \log\left(\frac{\exp(\langle \widehat{x}_i, \widehat{c}_j\rangle)}{\sum_{k=1}^n \exp(\langle \widehat{x}_i, \widehat{c}_k\rangle)}\right)$. If $\dim(\widehat{x}) = d + 1$ we can set one of the dimensions of $\widehat{x} = 1$ as a bias term allowing us to rewrite the objective with a slack parameter b_j analogously to `GloVe`. After reparametrization we obtain that for $\widehat{b} = b_j - ||\widehat{c}_j||_2^2$,

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

Since $C_{ij}/\sum_{k=1}^n C_{ik} \to \frac{\exp(-|||x_i - x_j||_2^2/\sigma^2)}{\sum_{k=1}^n \exp(-|||x_i - x_k||_2^2/\sigma^2)}$ this is the stochastic neighbor embedding (SNE) objective weighted by $\sum_{k=1}^n C_{ik}$. Once again, the true embedding $\widehat{x} = \widehat{c} = x/\sigma$ is one of the global minima (Theorem S1.5).

SVD: The SVD approach [10] factorizes the pointwise mutual information matrix. This case has analogous consistency results and is covered in [6].

3.1 Metric regression from log co-occurences

We have demonstrated that existing word embedding algorithms can be cast as metric recovery. However, it is not clear if this connection is coincidental. We propose a new model which directly models the log-linearity in equation 1 using generalized linear model, where the co-occurences C_{ij} follow a negative binomial distribution with mean $\exp(-||x_i - x_j||_2^2)$.

$$C_{ij} \sim \text{NegBin}\left(\theta, \theta(\theta + \exp(-||x_i - x_j||_2^2/2 + a_i + b_j))^{-1}\right).$$

The parameter θ controls the contribution of large C_{ij} and acts very similarly to `GloVe`'s $f(C_{ij})$ weight function. The advantage of this approach is that it combines the simplicity of optimization of `GloVe` without the choice of arbitrary weight function f. In our results we show that metric regression performs well at both word embedding and manifold learning.

4 Metric recovery from Markov processes on graphs and manifolds

We now substantially generalize the recovery conditions of the previous section by removing the Gaussian link between the metric and Markov transitions. We take an extreme view here and show that even a random walk over a sufficiently large *unweighted* directed graph holds enough information for metric recovery provided that the graph itself is suitably constructed in relation to the

[2] In practice, word embedding methods use a symmetrized window rather than counting transitions. This does not change any of the asymptotic analysis in the paper (Supplementary section S2 [6])

	Google (cos)			Google (L_2)			SAT		Classification		Sequence	
Method	Sem.	Synt.	Total	Sem.	Synt.	Total	L_2	Cosine	L_2	Cosine	L_2	Cosine
Regression	**78.4**	70.8	**73.7**	**75.5**	**70.9**	**72.6**	39.2	37.8	**87.6**	**84.6**	**58.3**	**59.0**
GloVE	72.6	71.2	71.7	65.6	66.6	67.2	36.9	33.6	73.1	80.1	48.8	**59.0**
SVD	57.4	50.8	53.4	53.7	48.2	50.3	27.1	25.8	65.2	74.6	52.4	53.0
Word2vec	73.4	**73.3**	73.3	71.4	**70.9**	71.1	**42.0**	**42.0**	76.4	**84.6**	54.4	56.2

Table 1: Accuracies on Google, SAT analogies and on two new verbal inductive tasks.

underlying metric.[3] To this end, we use a limiting argument (large vocabulary limit) with an increasing number of points $\mathcal{X}_n = \{x_1, \ldots, x_n\}$, where x_i are sampled i.i.d. from a density $p(x)$ on a manifold with geodesic ρ.

Definition 1 (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 ($p_{ij} = 0, 1$) include the ε ball graph ($\sigma_n(x) = \varepsilon$) and the k-nearest neighbor graph ($\sigma_n(x) =$distance to k-th neighbor).

Our main result (whose full details we defer to [6]) shows that co-occurence of random walks over graphs follows the same limit as the simple Gaussian random walk as above.

Theorem 2 (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:*

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

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

Theorem 2 proves the unversality of the log-linear limit $\log(C_{ij}/\sum_k C_{ik}) \to -||x_i - x_j||_2^2$, which extends the metric recovery properties of word embedding algorithms to any type of semantic random walk and justifies the ad-hoc methods which apply word embeddings to graphs [17].

5 Empirical validation

We experimentally validate our word embedding theory by training embeddings on 5.8B tokens combining Wikipedia with Gigaword5 emulating `GloVe`'s corpus. For results on other corpora, as well as implementation details for all methods, refer to the full version of this paper [6].

Solving analogies using survey data: We demonstrate that analogies can be solved simply by using human-generated semantic similarity scores. We take a free-association survey dataset [15], construct a graph with edge weights corresponding to log association frequency ($-\log(w_{ij}/\max_{kl}(w_{kl}))$) and embed this weighted graph using stochastic neighbor embedding (SNE) [7] and Isomap [21]. We then use these embeddings to solve questions from the Google analogy dataset [11]. Directly embedding semantic similarity with Isomap performed well (82.3% accuracy) and even outperformed corpus-based word embeddings obtained using word2vec (70.7%). Unsurprisingly, survey embeddings perform badly on the syntactic questions, as the survey was purely semantic.

Analogies: We then test our proposed method against other popular embedding schemes in the Google and SAT [23] analogy tasks. The results in Table 1 demonstrate that our proposed framework of metric regression and naive vector addition (L_2) is competitive with state-on-the-art embedding methods on this task. The performance gap across methods is small and fluctuates, but metric regression consistently outperforms all methods on semantic analogies and `GloVe` on most tasks.

Sequence and classification tasks: We propose two new difficult inductive reasoning tasks based upon the semantic field hypothesis [20]: series completion and classification. The questions were

[3]The weighted graph case follows identical arguments, see [6].

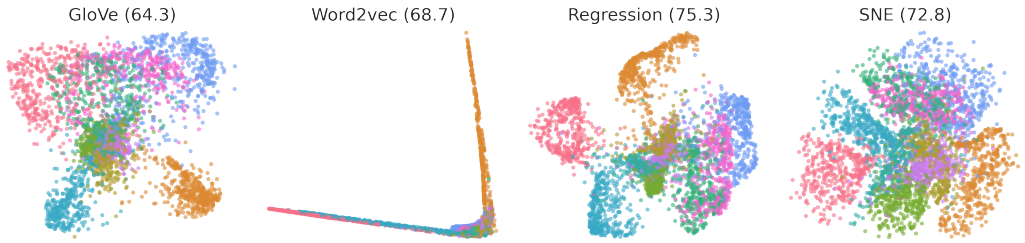

Figure 2: MNIST digit embedding using word embedding methods (left three) and metric embedding on the same graph (right). Performance is quantified by percentage of 5-nearest neighbors sharing the same cluster label.

generated using WordNet semantic relations [13]. Word embeddings solve both tasks effectively, with metric embedding consistently performing well on these multiple choice tasks (Table 1).

Manifold embedding the MNIST digits: Theorem 2 demonstrates that word embeddings can perform nonlinear dimensionality reduction. We test this by embedding the MNIST digits dataset [8]. Using a four-thousand point subset, we generated a k-nearest neighbor graph ($k = 20$) and generated 10 simple random walks of length 200 from each point. Treating these trajectories as sentences result in 40,000 sentences each of length 200. We compared the four word embedding methods against stochastic neighborhood embedding (SNE) on the percentage of 5-nearest neighbors sharing the same cluster label. Fig. 2 demonstrate that metric regression is highly effective at this task, outperforming metric SNE.

References

[1] S. Arora, Y. Li, Y. Liang, T. Ma, and A. Risteski. Random walks on context spaces: Towards an explanation of the mysteries of semantic word embeddings. *arXiv preprint arXiv:1502.03520*, 2015.

[2] J. Blitzer, A. Globerson, and F. Pereira. Distributed latent variable models of lexical co-occurrences. In *Proc. AISTATS*, pages 25–32, 2005.

[3] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[4] Y. Goldberg and O. Levy. word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method. *arXiv Prepr. arXiv1402.3722*, pages 1–5, 2014.

[5] Z. S. Harris. Distributional structure. *Word*, 1954.

[6] T. B. Hashimoto, D. Alvarez-Melis, and T. S. Jaakkola. Word, graph and manifold embedding from markov processes. *arXiv preprint arXiv:1509.05808*, 2015.

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

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[9] O. Levy and Y. Goldberg. Linguistic Regularities in Sparse and Explicit Word Representations. *Proc. 18th Conf. Comput. Nat. Lang. Learn. (CoNLL 2014)*, 2014.

[10] O. Levy and Y. Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, pages 2177–2185, 2014.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[13] G. Miller and C. Fellbaum. Wordnet: An electronic lexical database, 1998.

[14] A. Mnih and G. E. Hinton. A scalable hierarchical distributed language model. In *Advances in neural information processing systems*, pages 1081–1088, 2009.

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

[16] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12, 2014.

[17] B. Perozzi, R. Al-Rfou, and S. Skiena. 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, 2014.

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

[19] R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng. Parsing with compositional vector grammars. In *In Proceedings of the ACL conference*. Citeseer, 2013.

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

[21] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[22] J. Turian, L. Ratinov, and Y. Bengio. 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. Association for Computational Linguistics, 2010.

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