Capturing Nonlinear Structure in Word Spaces through Dimensionality Reduction

David Jurgens
University of California, Los Angeles,
4732 Boelter Hall
Los Angeles, CA 90095
jurgens@cs.ucla.edu

Keith Stevens
University of California, Los Angeles,
4732 Boelter Hall
Los Angeles, CA 90095
kstevens@cs.ucla.edu

Abstract

Dimensionality reduction has been shown to improve processing and information extraction from high dimensional data. Word space algorithms typically employ linear reduction techniques that assume the space is Euclidean. We investigate the effects of extracting nonlinear structure in the word space using Locality Preserving Projections, a reduction algorithm that performs manifold learning. We apply this reduction to two common word space models and show improved performance over the original models on benchmarks.

1 Introduction

Vector space models of semantics frequently employ some form of dimensionality reduction for improvement in representations or computational overhead. Many of the dimensionality reduction algorithms assume that the unreduced word space is linear. However, word similarities have been shown to exhibit many non-metric properties: asymmetry, e.g North Korea is more similar to Red China than Red China is to North Korea, and non-transitivity, e.g. Cuba is similar the former USSR, Jamaica is similar to Cuba, but Jamaica is not similar to the USSR (Tversky, 1977). We hypothesize that a non-linear word space model might more accurately preserve these non-metric relationships.

To test our hypothesis, we capture the nonlinear structure with dimensionality reduction by using Locality Preserving Projection (LPP) (He and Niyogi, 2003), an efficient, linear approximation of Eigenmaps (Belkin and Niyogi, 2002). With this reduction, the word space vectors are assumed to exist on a nonlinear manifold that LPP learns in order to project the vectors into a Euclidean space. We measure the effects of using LPP on two basic word space models: the Vector Space Model and a Word Co-occurrence model. We begin with a brief overview of these word spaces and common dimensionality reduction techniques. We then formally introduce LPP. Following, we use two experiments to demonstrate LPP's capacity to accurately dimensionally reduce word spaces.

2 Word Spaces and Reductions

We consider two common word space models that have been used with dimensionality reduction. The first is the Vector Space Model (VSM) (Salton et al., 1975). Words are represented as vectors where each dimension corresponds to a document in the corpus and the dimension's value is the number of times the word occurred in the document. We label the second model the Word Co-occurrence (WC) model: each dimension correspond to a unique word, with the dimension's value indicating the number of times that dimension's word co-occurred.

Dimensionality reduction has been applied to both models for three kinds of benefits: to improve computational efficiency, to capture higher order relationships between words, and to reduce noise by smoothing or eliminating noisy features. We consider three of the most popular reduction techniques and the general word space models to which they have been applied: linear projections, feature elimination and random approximations.

The most frequently applied linear projection technique is the Singular Value Decomposition (SVD). The SVD factors a matrix A, which represents a word space, into three matrices $U\Sigma V^\top$ such that Σ is a diagonal matrix containing the singular values of A, ordered descending based on their effect on the variance in the values of A. The original matrix can be approximated by using only the top k singular values, setting all others to 0. The approximation matrix, $\hat{A} = U_k\Sigma_k V_k^\top$, is the least squares best-fit rank-k approximation of A.

Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics, ACL 2010, pages 1–6,
Uppsala, Sweden, 16 July 2010. ©2010 Association for Computational Linguistics

The SVD has been used with great success on both models. Latent Semantic Analysis (LSA) (Landauer et al., 1998) extends the (VSM) by decomposing the space using the SVD and making the word space the left singular vectors, U_k. WC models have also utilized the SVD to improve performance (Schütze, 1992; Bullinaria and Levy, 2007; Baroni and Lenci, 2008).

Feature elimination reduces the dimensionality by removing those with low information content. This approach has been successfully applied to WC models such as HAL (Lund and Burgess, 1996) by dropping those with low entropy. This technique effectively removes the feature dimensions of high frequency words, which provide little discriminatory content.

Randomized projections have also been successfully applied to VSM models, e.g. (Kanerva et al., 2000) and WC models, e.g. (Sahlgren et al., 2008). This reduction statistically approximates the original space in a much lower dimensional space. The projection does not take into account the structure of data, which provides only a computational benefit from fewer dimensions, unlike the previous two reductions.

3 Locality Preserving Projection

For a set of vectors, $x_1, x_2, \ldots, x_n \in \mathbb{R}^m$, LPP preserves the distance in the k-dimensional space, where $k \ll m$, by solving the following minimization problem,

$$\min_{w} \sum_{ij} (\mathbf{w}^\top \mathbf{x}_i - \mathbf{w}^\top \mathbf{x}_j)^2 S_{ij} \tag{1}$$

where $\mathbf{w}$ is a transformation vector that projects $\mathbf{x}$ into the lower dimensional space, and S is a matrix that represents the local structure of the original space. Minimizing this equation is equivalent to finding the transformation vector that best preserves the local distances in the original space according to S. LPP assumes that the data points $\mathbf{x}_i$ exist on a manifold. This is in contrast to the SVD, which assumes that the space is Euclidean and performs a global, rather than local, minimization. In treating the space as a manifold, LPP is able to discover some of the nonlinear structure of the data from its local structure.

To solve the minimization problem in Equation 1, LPP uses a linear approximation of the Laplacian Eigenmaps procedure (Belkin and Niyogi, 2002) as follows:

1. Let X be a matrix where $\mathbf{x}_i$ is the i^{th} row vector. Construct an adjacency matrix, S, which represents the local structure of the original vector space, by making an edge between points $\mathbf{x}_i$ and $\mathbf{x}_j$ if $\mathbf{x}_j$ is locally proximate to $\mathbf{x}_i$. Two variations are available for determining proximity: either the k-nearest neighbors, or all the data points with similarity $> \epsilon$.
2. Weight the edges in S proportional to the closeness of the data points. Four main options are available: a Gaussian kernel, a polynomial kernel, cosine similarity, or binary.
3. Construct the diagonal matrix D where entry $D_{ii} = \sum_j S_{ij}$. Let $L = D - S$. Then solve the generalized eigenvector problem:
$$XLX^\top \mathbf{w} = \lambda XDX^\top \mathbf{w}. \tag{2}$$
He and Niyogi (2003) show that solving this problem is equivalent to solving Equation 1.
4. Let W_k = $[\mathbf{w}_1, \ldots, \mathbf{w}_k]$ denote the matrix of transformation vectors, sorted in descending order according to their eigenvalues λ. The original space is projected into k dimensions by $W_k^\top X \rightarrow X_k$.

For many applications of LPP, such as document clustering (He et al., 2004), the original data matrix X is transformed by first performing Principle Component Analysis and discarding the smallest principle components, which requires computing the full SVD. However, for large data sets such as those frequently used in word space algorithms, performing the full SVD is computationally infeasible.

To overcome this limitation, Cai et al. (2007a) show how Spectral Regression may be used as an alternative for solving the same minimization equation through an iterative process. The principle idea is that Equation 2 may be recast as

$$S\mathbf{y} = \lambda D\mathbf{y} \tag{3}$$

where $\mathbf{y} = X^\top \mathbf{w}$, which ensures $\mathbf{y}$ will be an eigenvector with the same eigenvalue for the problem in Equation 2. Finding the transformation matrix W_k, used in step 4, is done in two steps. First, Equation 3 is solved to produce eigenvectors $[\mathbf{y}_0, \ldots, \mathbf{y}_k]$, sorted in decreasing order according to their eigenvalues λ. Second, the set of transformation vectors composing W_k, $[\mathbf{w}_1, \ldots, \mathbf{w}_k]$, is found by a least-squares regression:

$$\mathbf{w}_j = \operatorname*{argmin}_{\mathbf{w}} \sum_{i=1}^{n} (\mathbf{w}^\top \mathbf{x}_i - \mathbf{y}_i^j)^2 + \alpha ||\mathbf{w}||^2 \tag{4}$$

where $\mathbf{y}_i^j$ denotes the value of the j^{th} dimension of $\mathbf{y}_i$. The α parameter penalizes solutions proportionally to their magnitude, which Cai et al. (2007b) note ensures the stability of $\mathbf{w}$ as an approximate eigenproblem solution.

4 Experiments

Two experiments measures the effects of nonlinear dimensionality reduction for word spaces. For both, we apply LPP to two basic word space models, the VSM and WC. In the first experiment, we measure the word spaces' abilities to model semantic relations, as determined by priming experiments. In the second experiment, we evaluate the representation capabilities of the LPP-reduced models on standard word space benchmarks.

4.1 Setup

For the VSM-based word space, we consider three different weighting schemes: no weighting, TF-IDF and the log-entropy (LE) used in (Landauer et al., 1998). For the WC-based word space, we use a 5 word sliding window. Due to the large parameter space for LPP models, we performed only a limited configuration search. An initial analysis using the 20 nearest neighbors and cosine similarity did not show significant performance differences when the number of dimensions was varied between 50 and 1000. We therefore selected 300 dimensions for all tests. Further work is needed to identify the impact of different parameters. Stop words were removed only for the WC+LPP model. We compare the LPP-based spaces to three models: VSM, HAL, and LSA.

Two corpora are used to train the models in both experiments. The first corpus, TASA, is a collection of 44,486 essays that are representative of the reading a student might see upon entering college, introduced by (Landauer et al., 1998). The corpus consists of 98,420 unique words; no filtering is done when processing this corpus. The second corpus, WIKI, is a 387,082 article subset of a December 2009 Wikipedia snapshot consisting of all the articles with more than 1,000 tokens. The corpus is filtered to retain the top 100,000 most frequent tokens in addition to all the tokens used in each experiment's data set.

4.2 Experiment 1

Semantic priming measures word association based on human responses to a provided cue. Priming studies have been used to evaluate word spaces by equating vector similarity with an increased priming response. We use data from two types of priming experiments to measure whether LPP models better correlate with human performance than non-LPP word spaces.

Normed Priming Nelson et al. (1998) collected free association responses to 5,019 prime words. An average of 149 participants responded to each prime with the first word that came to mind.

Based on this dataset, we introduce a new benchmark that correlates word space similarity with the associative strength of semantic priming pairs. We use three measures for modeling prime-target strength, which were inspired by Steyvers et al. (2004). Let W_{ab} be the percentage of participants who responded to prime a with target b. The three measures of associative strength are

$$\begin{aligned} S^1_{ab} &= W_{ab} \\ S^2_{ab} &= W_{ab} + W_{ba} \\ S^3_{ab} &= S^2_{ab} + \textstyle\sum_c S^2_{ac} S^2_{cb} \end{aligned}$$

These measure three different levels of semantic relatedness between words a and b. S^1_{ab} measures the relationship from a to b, which is frequently asymmetric due to ordering, e.g. "orange" produces "juice" more frequently than "juice" produces "orange." S^2_{ab} measures the symmetric association between a and b; Steyvers et al. (2004) note that this may better model the associative strength by including weaker associates that may have been a suitable second response. S^3_{ab} further increases the association by including the indirect associations between a and b from all cued primes.

For each measure, we rank a prime's targets according to their strength and then compute the Spearman rank correlation with the prime-target similarities in the word space. The rank comparison measures how well word space similarity corresponds to the priming association. We report the average rank correlation of associational strengths over all primes.

Priming Effect The priming study by Hodgson (1991), which evaluated how different semantic relationships affected the strength of priming, provides the data for our second priming test. Six relationships were examined in the study: antonymy, synonymy, conceptual association (sleep and bed), categorical coordinates (mist and rain), phrasal associates (pony and express), and super- and subordinates. Each relationship contained an average

	Antonymy			Conceptual			Coordinates		
Algorithm	R[b]	U	E	R	U	E	R	U	E
VSM+LPP+LE	0.103	0.018	0.085	0.197	0.050	0.147	0.071	0.027	0.044
VSM+LPP+TF-IDF	0.348	0.321	0.027	0.408	0.414	-0.005	0.323	0.294	0.029
VSM+LPP	0.247	0.122	0.124	0.312	0.120	0.193	0.230	0.111	0.119
VSM+LPP[a]	0.298	0.070	**0.228**	0.284	0.033	0.252	0.321	0.037	**0.284**
WC+LPP	0.255	0.071	0.185	0.413	0.110	0.303	0.431	0.134	**0.298**
HAL	0.813	0.716	0.096	0.845	0.814	0.031	0.861	0.809	0.052
HAL[a]	0.915	0.879	0.037	0.867	0.846	0.021	0.913	0.861	0.052
LSA	0.235	0.023	0.213	0.392	0.028	**0.364**	0.199	0.014	0.185
LSA[a]	0.287	0.061	0.226	0.362	0.041	0.321	0.316	0.037	0.278
VSM	0.051	0.011	0.040	0.111	0.012	0.099	0.032	0.008	0.024

	Phrasal			Ordinates			Synonymy		
Algorithm	R	U	E	R	U	E	R	U	E
VSM+LPP+LE	0.147	0.039	0.108	0.225	0.032	0.193	0.081	0.027	0.053
VSM+LPP+TF-IDF	0.438	0.425	0.013	0.277	0.290	-0.013	0.344	0.328	0.017
VSM+LPP	0.234	0.107	0.127	0.273	0.115	0.158	0.237	0.157	0.080
VSM+LPP[a]	0.202	0.031	0.171	0.270	0.032	0.238	0.299	0.069	0.230
WC+LPP	0.274	0.087	0.186	0.324	0.076	0.248	0.345	0.111	0.233
HAL	0.805	0.776	0.029	0.825	0.789	0.036	0.757	0.681	0.076
HAL[a]	0.866	0.856	0.010	0.881	0.857	0.024	0.898	0.879	0.019
LSA	0.280	0.021	**0.258**	0.258	0.018	0.240	0.197	0.019	0.178
LSA[a]	0.269	0.030	0.238	0.326	0.032	**0.294**	0.327	0.052	**0.275**
VSM	0.104	0.013	0.091	0.061	0.008	0.053	0.052	0.009	0.043

[a] Processed using the WIKI corpus

[b] R are related primes, U are unrelated primes, E is the priming effect

Table 1: Experiment 1 priming results for the six relation categories from Hodgson (1991)

		Word Choice			Word Association		
Algorithm	Corpus	TOEFL	ESL	RDWP	F. et al.	R.&G.	Deese
VSM+LPP+le	TASA	24.000	50.000	45.313	0.296	0.092	0.034
VSM+LPP+tf-idf	TASA	22.667	25.000	37.209	0.023	0.086	0.001
VSM+LPP	TASA	41.333	**54.167**	39.063	0.219	0.136	0.045
VSM+LPP	Wiki	33.898	48.780	43.434	0.530	0.503	0.108
WC+LPP	TASA	46.032	40.000	45.783	0.423	0.414	0.126
HAL	TASA	44.00	20.83	50.00	0.173	0.180	0.318
HAL	Wiki	50.00	31.11	43.44	0.261	0.195	0.042
LSA	TASA	56.000	50.000	55.814	0.516	0.651	**0.349**
LSA	Wiki	60.759	**54.167**	59.200	**0.614**	**0.681**	0.206
VSM	TASA	**61.333**	52.083	**84.884**	0.396	0.496	0.200

Table 2: Results from Experiment 2 on six word space benchmarks

of 23 word pairs. Hodgson's results showed that priming effects were exhibited by the prime-target pairs in all six categories.

We use the same methodology as Padó and Lapata (2007) for this data set; the prime-target (Related Primes) cosine similarity is compared with the average cosine similarity between the prime and all other targets (Unrelated Primes) within the semantic category. The priming effect is the difference between the two similarity values.

4.3 Experiment 2

We use six standard word space benchmarks to test our hypothesis that LPP can accurately capture general semantic knowledge and association based relations. The benchmarks come in two forms: word association and word choice tests.

Word choice tests provide a target word and a list of options, one of which has the desired relation to the target. To answer these questions, we select the option with the highest cosine similarity with the target. Three word choice synonymy benchmarks are used: the Test of English as a Foreign Language (TOEFL) test set from (Landauer et al., 1998), the English as a Second Language (ESL) test set from (Turney, 2001), and the Canadian Reader's Digest Word Power (RDWP) from (Jarmasz and Szpakowicz, 2003).

Algorithm	Corpus	S^1	S^2	S^3
VSM+LPP+LE	TASA	0.457	0.413	0.255
VSM+LPP+TF-IDF	TASA	0.464	0.390	0.207
VSM+LPP	TASA	0.457	0.427	0.275
VSM+LPP	Wiki	0.472	0.440	0.333
WC+LPP	TASA	0.469	0.437	0.315
HAL	TASA	0.485	0.434	0.310
HAL	Wiki	0.462	0.406	0.266
LSA	TASA	**0.494**	**0.481**	**0.414**
LSA	Wiki	0.489	0.472	0.398
VSM	TASA	0.484	0.460	0.407

Table 3: Experiment 1 results for normed priming.

Word association tests measure the semantic relatedness of two words by comparing their similarity in the word space with human judgements. These tests are more precise than word choice tests because they take into account the specific value of the word similarity. Three word association benchmarks are used: the word similarity data set of Rubenstein and Goodenough (1965), the word-relatedness data set of Finkelstein et al. (2002), and the antonymy data set of Deese (1964), which measures the degree to which high similarity captures the antonymy relationship. The Finkelstein et al. test is notable in that the human judges were free to score based on any word relationship.

5 Results and Discussion

The LPP-based models show mixed performance in comparison to existing models on normed priming tasks, shown in Table 3. Adding LPP to the VSM decreased performance; however, when WIKI was used instead of TASA, the VSM+LPP model increased .15 on all correlations, whereas LSA's performance decreased. This suggests that LPP needs more data than LSA to properly model the word space manifold. WC+LPP performs comparably to HAL, which indicates that LPP is effective in retaining the original WC space's structure in significantly fewer dimensions.

For the categorical priming tests shown in Table 1, LPP-based models show competitive results. VSM+LPP with the WIKI corpus performs much better than other VSM+LPP configurations. Unlike in the previous priming experiment, adding LPP to the base models resulted in a significant performance improvement. We also note that both HAL models and the VSM+LPP+TF-IDF model have high similarity ratings for unrelated primes. We posit that these models' feature weighting results in poor differentiation between words in the same semantic category, which causes their decreased performance.

For experiment 2, LPP-based spaces showed mixed results on word choice benchmarks, while showing notable improvement on the more precise word association benchmarks. Table 2 lists the results. Notably, LPP-based spaces performed well on the ESL synonym benchmark but poorly on the TOEFL synonym benchmark, even when the larger WIKI corpus was used. This suggests that LPP was not effective in retaining the relationship between certain classes of synonyms. Given that performance did not improve with the WIKI corpus, further analysis is needed to identify whether a different representation of the local structure would improve results or if the poor performance is due to another factor. While LSA and VSM model performed best on all benchmarks, LPP-based spaces performed competitively on the word association tests. In all but two tests, the WC+LPP model outperformed HAL.

The results from both experiments indicate that LPP is capable of accurately representing distributional information in a much lower dimensional space. However, in many cases, applications using the SVD-reduced representations performed better. In addition, application of standard weighting schemes worsened LPP-models' performance, which suggests that the local neighborhood is adversely distorted. Nevertheless, we view these results as a promising starting point for further evaluation of nonlinear dimensionality reduction.

6 Conclusions and Future Work

We have shown that LPP is an effective dimensionality reduction technique for word space algorithms. In several benchmarks, LPP provided a significant benefit to the base models and in a few cases outperformed the SVD. However, it does not perform consistently better than existing models. Future work will focus on four themes: identifying optimal LPP parameter configurations; improving LPP with weighting; measuring LPP's capacity to capture higher order co-occurrence relationships, as was shown for the SVD (Lemaire et al., 2006); and investigating whether more computationally expensive nonlinear reduction algorithms such as ISOMAP (Tenenbaum et al., 2000) are better for word space algorithms. We plan to release implementations of the LPP-based models as a part of the S-Space Package (Jurgens and Stevens, 2010).

References

Marco Baroni and Alessandro Lenci. 2008. Concepts and properties in word spaces. *From context to meaning: Distributional models of the lexicon in linguistics and cognitive science (Special issue of the Italian Journal of Linguistics)*, 1(20):55–88.

Mikhail Belkin and Partha Niyogi. 2002. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. In *Advances in Neural Information Processing Systems*, number 14.

John A. Bullinaria and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: a computational study. *Behavior Research Methods*, 39:510–526.

Deng Cai, Xiaofei He, and Jiawei Han. 2007a. Spectral regression for efficient regularized subspace learning. In *IEEE International Conference on Computer Vision (ICCV'07)*.

Deng Cai, Xiaofei He, Wei Vivian Zhang, , and Jiawei Han. 2007b. Regularized Locality Preserving Indexing via Spectral Regression. In *Proceedings of the 2007 ACM International Conference on Information and Knowledge Management (CIKM'07)*.

James Deese. 1964. The associative structure of some common english adjectives. *Journal of Verbal Learning and Verbal Behavior*, 3(5):347–357.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Woflman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions of Information Systems*, 20(1):116–131.

Xiaofei He and Partha Niyogi. 2003. Locality preserving projections. In *Advances in Neural Information Processing Systems 16 (NIPS)*.

Xiaofei He, Deng Cai, Haifeng Liu, and Wei-Ying Ma. 2004. Locality preserving indexing for document representation. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 96–103.

James M. Hodgson. 1991. Informational constraints on pre-lexical priming. *Language and Cognitive Processes*, 6:169–205.

Mario Jarmasz and Stan Szpakowicz. 2003. Roget's thesaurus and semantic similarity. In *Conference on Recent Advances in Natural Language Processing*, pages 212–219.

David Jurgens and Keith Stevens. 2010. The S-Space Package: An Open Source Package for Word Space Models. In *Proceedings of the ACL 2010 System Demonstrations*.

Pentti Kanerva, Jan Kristoferson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In L. R. Gleitman and A. K. Josh, editors, *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, page 1036.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. Introduction to Latent Semantic Analysis. *Discourse Processes*, (25):259–284.

Benoît Lemaire, , and Guy Henhiére. 2006. Effects of High-Order Co-occurrences on Word Semantic Similarities. *Current Psychology Letters*, 1(18).

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occrrence. *Behavoir Research Methods, Instruments & Computers*, 28(2):203–208.

Douglas L. Nelson, Cathy L. McEvoy, and Thomas A. Schreiber. 1998. The University of South Florida word association, rhyme, and word fragment norms. `http://www.usf.edu/FreeAssociation/`.

Sebastian Padó and Mirella Lapata. 2007. Dependency-Based Construction of Semantic Space Models. *Computational Linguistics*, 33(2):161–199.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8:627–633.

Magnus Sahlgren, Anders Holst, and Pentti Kanerva. 2008. Permutations as a means to encode order in word space. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society (CogSci'08)*.

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

Hinrich Schütze. 1992. Dimensions of meaning. In *Proceedings of Supercomputing '92*, pages 787–796.

Mark Steyvers, Richard M. Shiffrin, and Douglas L. Nelson, 2004. *Word association spaces for predicting semantic similarity effects in episodic memory*. American Psychological Assocation.

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

Peter D. Turney. 2001. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*, pages 491–502.

Amos Tversky. 1977. Features of similarity. *Psychological Review*, 84:327–352.