Multiview LSA: Representation Learning via Generalized CCA

Pushpendre Rastogi[1] and **Benjamin Van Durme**[1,2] and **Raman Arora**[1]
[1]Center for Language and Speech Processing
[2]Human Language Technology Center of Excellence
Johns Hopkins University

Abstract

Multiview LSA (MVLSA) is a generalization of Latent Semantic Analysis (LSA) that supports the fusion of arbitrary views of data and relies on Generalized Canonical Correlation Analysis (GCCA). We present an algorithm for fast approximate computation of GCCA, which when coupled with methods for handling missing values, is general enough to approximate some recent algorithms for inducing vector representations of words. Experiments across a comprehensive collection of test-sets show our approach to be competitive with the state of the art.

1 Introduction

Winograd (1972) wrote that: *"Two sentences are paraphrases if they produce the same representation in the internal formalism for meaning"*. This intuition is made soft in vector-space models (Turney and Pantel, 2010), where we say that expressions in language are paraphrases if their representations are *close* under some distance measure.

One of the earliest linguistic vector space models was Latent Semantic Analysis (LSA). LSA has been successfully used for Information Retrieval but it is limited in its reliance on a single matrix, or *view*, of term co-occurrences. Here we address the single-view limitation of LSA by demonstrating that the framework of Generalized Canonical Correlation Analysis (GCCA) can be used to perform Multiview LSA (MVLSA). This approach allows for the use of an arbitrary number of views in the induction process, including embeddings induced using other algorithms. We also present a fast approximate method for performing GCCA and approximately recover the objective of (Pennington et al., 2014) while accounting for missing values.

Our experiments show that MVLSA is competitive with state of the art approached for inducing vector representations of words and phrases. As a methodological aside, we discuss the (in-)significance of conclusions being drawn from comparisons done on small sized datasets.

2 Motivation

LSA is an application of Principal Component Analysis (PCA) to a term-document cooccurrence matrix. The principal directions found by PCA form the basis of the vector-space in which to represent the input terms (Landauer and Dumais, 1997). A drawback of PCA is that it can leverage only a single source of data and it is sensitive to scaling.

An arguably better approach to representation learning is Canonical Correlation Analysis (CCA) that induces representations that are maximally *correlated* across two views, allowing the utilization of two distinct sources of data. While an improvement over PCA, being limited to only two views is unfortunate in light of the fact that many sources of data (perspectives) are frequently available in practice. In such cases it is natural to extend CCA's original objective of maximizing correlation between two views by maximizing some measure of the matrix Φ that contains all the pairwise correlations between linear projections of the *covariates*. This is how Generalized Canonical Correlation Analysis (GCCA) was first derived by Horst (1961). Recently these intuitive ideas about benefits of leveraging multiple sources of data have received strong theoretical backing due to the work by Sridharan and

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 556–566,
Denver, Colorado, May 31 – June 5, 2015.

Kakade (2008) who showed that learning with multiple views is beneficial since it reduces the complexity of the learning problem by restricting the search space. Recent work by Anandkumar et al. (2014) showed that at least three views are necessary for recovering hidden variable models.

Note that there exist different variants of GCCA depending on the measure of Φ that we choose to maximize. Kettenring (1971) enumerated a variety of possible measures, such as the spectral-norm of Φ. Kettenring noted that maximizing this spectral-norm is equivalent to finding linear projections of the *covariates* that are most amenable to rank-one PCA, or that can be best explained by a single term factor model. This variant was named *MAX-VAR GCCA* and was shown to be equivalent to a proposal by Carroll (1968), which searched for an auxiliary orthogonal representation G that was maximally correlated to the linear projections of the covariates. Carroll's objective targets the intuition that representations leveraging multiple views should correlate with all provided views as much as possible.

3 Proposed Method: MVLSA

Let $X_j \in \mathbb{R}^{N \times d_j}\ \forall j \in [1, \ldots, J]$ be the mean centered matrix containing data from view j such that row i of X_j contains the information for word w_i. Let the number of words in the vocabulary be N and number of contexts (columns in X_j) be d_j. Following standard notation (Hastie et al., 2009) we call $X_j^\top X_j$ the scatter matrix and $X_j(X_j^\top X_j)^{-1}X_j^\top$ the projection matrix.

The objective of *MAX-VAR GCCA* can be written as the following optimization problem: Find $G \in \mathbb{R}^{N \times r}$ and $U_j \in \mathbb{R}^{d_j \times r}$ that solve:

$$\begin{aligned} &\underset{G,U_j}{\arg\min} \sum_{j=1}^{J} \left\| G - X_j U_j \right\|_F^2 \\ &\text{subject to } G^\top G = I. \end{aligned} \tag{1}$$

The matrix G that solves problem (1) is our vector representation of the vocabulary. Finding G reduces to spectral decomposition of sum of projection matrices of different views: Define

$$P_j = X_j(X_j^\top X_j)^{-1}X_j^\top, \tag{2}$$

$$M = \sum_{j=1}^{J} P_j. \tag{3}$$

Then, for some positive diagonal matrix Λ, G and U_j satisfy:

$$MG = G\Lambda, \tag{4}$$

$$U_j = \left(X_j^\top X_j\right)^{-1} X_j^\top G. \tag{5}$$

Computationally storing $P_j \in \mathbb{R}^{N \times N}$ is problematic owing to memory constraints. Further, the scatter matrices may be non-singular leading to an ill-posed procedure. We now describe a novel scalable GCCA with ℓ_2-regularization to address these issues.

Approximate Regularized GCCA: GCCA can be regularized by adding $r_j I$ to scatter matrix $X_j^\top X_j$ before doing the inversion where r_j is a small constant e.g. 10^{-8}. Projection matrices in (2) and (3) can then be written as

$$\widetilde{P}_j = X_j(X_j^\top X_j + r_j I)^{-1}X_j^\top, \tag{6}$$

$$M = \sum_{j=1}^{J} \widetilde{P}_j. \tag{7}$$

Next, to scale up GCCA to large datasets, we first form a rank-m approximation of projection matrices (Arora and Livescu, 2012) and then extend it to an eigendecomposition for M following ideas by Savostyanov (2014). Consider the rank-m SVD of X_j:

$$X_j = A_j S_j B_j^\top,$$

where $S_j \in \mathbb{R}^{m \times m}$ is the diagonal matrix with m-largest singular values of X_j and $A_j \in \mathbb{R}^{N \times m}$ and $B_j \in \mathbb{R}^{m \times d_j}$ are the corresponding left and right singular vectors. Given this SVD, write the j^{th} projection matrix as

$$\begin{aligned} \widetilde{P}_j &= A_j S_j^\top (r_j I + S_j S_j^\top)^{-1} S_j A_j^\top, \\ &= A_j T_j T_j^\top A_j^\top, \end{aligned}$$

where $T_j \in \mathbb{R}^{m \times m}$ is a diagonal matrix such that $T_j T_j^\top = S_j^\top (r_j I + S_j S_j^\top)^{-1} S_j$. Finally, we note

that the sum of projection matrices can be expressed as $M = \tilde{M}\tilde{M}^\top$ where

$$\tilde{M} = [A_1 T_1 \dots A_J T_J] \in \mathbb{R}^{N \times mJ}.$$

Therefore, eigenvectors of matrix M, i.e. the matrix G that we are interested in finding, are the left singular vectors of $\tilde{M}$, i.e. $\tilde{M} = GSV^\top$. These left singular vectors can be computed by using Incremental PCA (Brand, 2002) since $\tilde{M}$ may be too large to fit in memory.

3.1 Computing SVD of mean centered X_j

Recall that we assumed X_j to be mean centered matrices. Let $Z_j \in \mathbb{R}^{N \times d_j}$ be sparse matrices containing mean-uncentered cooccurrence counts. Let $f_j = n_j \circ t_j$ be the preprocessing function that we apply to Z_j:

$$Y_j = f_j(Z_j), \tag{8}$$
$$X_j = Y_j - 1(1^\top Y_j). \tag{9}$$

In order to compute the SVD of mean centered matrices X_j we first compute the partial SVD of uncentered matrix Y_j and then update it (Brand (2006) provides details). We experimented with representations created from the uncentered matrices Y_j and found that they performed as well as the mean centered versions but we would not mention them further since it is computationally efficient to follow the principled approach. We note, however, that even the method of mean-centering the SVD produces an approximation.

3.2 Handling missing rows across views

With real data it may happen that a term was not observed in a view at all. A large number of missing rows can corrupt the learnt representations since the rows in the left singular matrix become zero. To counter this problem we adopt a variant of the "missing-data passive" algorithm from Van De Velden and Bijmolt (2006) who modified the GCCA objective to counter the problem of missing rows.[1] The objective now becomes:

$$\begin{aligned} &\arg\min_{G,U_j} \sum_{j=1}^{J} \left\| K_j(G - X_j U_j) \right\|_F^2 \\ &\text{subject to } G^\top G = I, \end{aligned} \tag{10}$$

where $[K_j]_{ii} = 1$ if row i of view j is observed and zero otherwise. Essentially K_j is a diagonal row-selection matrix which ensures that we optimize our representations only on the observed rows. Note that $X_j = K_j X_j$ since the rows that K_j removed were already zero. Let, $K = \sum_j K_j$ then the optima of the objective can be computed by modifying equation (7) as:

$$M = K^{-\frac{1}{2}} (\sum_{j=1}^{J} P_j) K^{-\frac{1}{2}}. \tag{11}$$

Again, if we regularize and approximate the GCCA solution we get G as the left singular vectors of $K^{-\frac{1}{2}}\tilde{M}$. We mean center the matrices using only the observed rows.

Also note that other heuristic weighting schemes could be used here. For example if we modify our objective as follows then we would approximately recover the objective of Pennington et al. (2014):

$$\begin{aligned} &\underset{G,U_j}{\text{minimize:}} \quad \sum_{j=1}^{J} \left\| W_j K_j(G - X_j U_j) \right\|_F^2 \\ &\text{subject to:} \quad G^\top G = I \end{aligned} \tag{12}$$

where

$$[W_j]_{ii} = \left(\frac{w_i}{w_{\max}} \right)^{\frac{3}{4}} \text{ if } w_i < w_{\max} \text{ else } 1,$$
$$\text{and } w_i = \sum_k [X_j]_{ik}.$$

4 Data

Training Data We used the English portion of the *Polyglot* Wikipedia dataset released by Al-Rfou et

[1] A more recent effort, by van de Velden and Takane (2012), describes newer iterative and non-iterative (Test-Equating Method) approaches for handling missing values. It is possible that using one of those methods could improve performance.

al. (2013) to create 15 *irredundant* views of cooccurrence statistics where element $[z]_{ij}$ of view Z_k represents that number of times word w_j occurred k words behind w_i. We selected the top 500K words by occurrence to create our vocabulary for the rest of the paper.

We extracted cooccurrence statistics from a large bitext corpus that was made by combining a number of parallel bilingual corpora as part of the ParaPhrase DataBase (PPDB) project: Table 1 gives a summary, Ganitkevitch et al. (2013) provides further details. Element $[z]_{ij}$ of the *bitext* matrix represents the number of times English word w_i was automatically aligned to the foreign word w_j.

We also used the dependency relations in the *Annotated Gigaword Corpus* (Napoles et al., 2012) to create 21 views[2] where element $[z]_{ij}$ of view Z_d represents the number of times word w_j occurred as the governor of word w_i under dependency relation d.

We combined the knowledge of paraphrases present in FrameNet and PPDB by using the dataset created by Rastogi and Van Durme (2014) to construct a *FrameNet* view. Element $[z]_{ij}$ of the *FrameNet* view represents whether word w_i was present in frame f_j. Similarly we combined the knowledge of morphology present in the *CatVar* database released by Habash and Dorr (2003) and *morpha* released by Minnen et al. (2001) along with *morphy* that is a part of WordNet. The morphological views and the frame semantic views were especially sparse with densities of 0.0003% and 0.03%. While the approach allows for an arbitrary number of distinct sources of semantic information, such as going further to include cooccurrence in WordNet synsets, we considered the described views to be representative, with further improvements possible as future work.

Test Data We evaluated the representations on the word similarity datasets listed in Table 2. The first 10 datasets in Table 2 were annotated with different rubrics and rated on different scales. But broadly they all contain human judgements about how similar two words are. The "AN-SYN" and "AN-SEM" datasets contain 4-tuples of analogous words and the task is to predict the missing word given the first three. Both of these are open vocabulary tasks while TOEFL is a closed vocabulary task.

[2]Dependency relations employed: nsubj, amod, advmod, rcmod, dobj, prep_of, prep_in, prep_to, prep_on, prep_for, prep_with, prep_from, prep_at, prep_by, prep_as, prep_between, xsubj, agent, conj_and, conj_but, pobj.

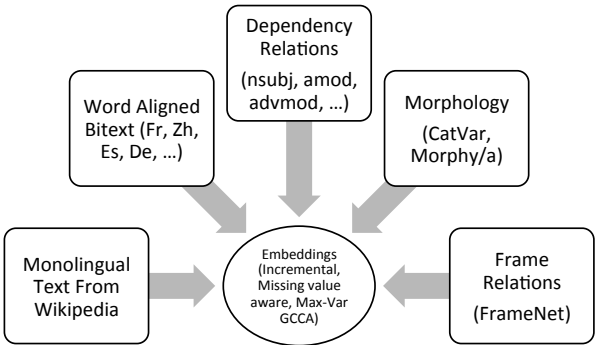

Figure 1: An illustration of datasets used.

Language	Sentences	English Tokens
Bitext-Arabic	8.8M	190M
Bitext-Czech	7.3M	17M
Bitext-German	1.8M	44M
Bitext-Spanish	11.1M	241M
Bitext-French	30.9M	671M
Bitext-Chinese	10.3M	215M
Monotext-En-Wiki	75M	1700M

Table 1: Portion of data used to create GCCA representations (in millions).

4.1 Significance of comparison

While surveying the literature we found that performance on word similarity datasets is typically reported in terms of the Spearman correlation between the gold ratings and the cosine distance between normalized embeddings. However researchers do not report measures of significance of the difference between the Spearman Correlations even for comparisons on small evaluation sets.[3] This motivated our defining a method for calculating the *Minimum Required Difference for Significance (MRDS)*.

Minimum Required Difference for Significance (MRDS): Imagine two lists of ratings over the same

[3]For example, the comparative difference by competing algorithms reported by Faruqui et al. (2014) could not be significant for the Word Similarity test set released by Finkelstein et al. (2001), even if we assumed a correlation between competing methods as high as 0.9, with a p value threshold of 0.05. Similar such comparisons on small datasets are performed by Hill et al. (2014a).

Acronym	Size	$\sigma^{0.5}_{0.01}$	$\sigma^{0.7}_{0.01}$	$\sigma^{0.9}_{0.01}$	$\sigma^{0.5}_{0.05}$	$\sigma^{0.7}_{0.05}$	$\sigma^{0.9}_{0.05}$	Reference
MEN	3000	4.2	3.2	1.8	3.0	2.3	1.3	(Bruni et al., 2012)
RW	2034	5.1	3.9	2.3	3.6	2.8	1.6	(Luong et al., 2013)
SCWS	2003	5.1	4.0	2.3	3.6	2.8	1.6	(Huang et al., 2012)
SIMLEX	999	7.3	5.7	3.2	5.2	4.0	2.3	(Hill et al., 2014b)
WS	353	12.3	9.5	5.5	8.7	6.7	3.9	(Finkelstein et al., 2001)
MTURK	287	13.7	10.6	6.1	9.7	7.5	4.3	(Radinsky et al., 2011)
WS-REL	252	14.6	11.3	6.5	10.3	8.0	4.6	(Agirre et al., 2009)
WS-SEM	203	16.2	12.6	7.3	11.5	8.9	5.1	-Same-As-Above-
RG	65	28.6	22.3	12.9	20.6	16.0	9.2	(Rubenstein and Goodenough, 1965)
MC	30	41.7	32.7	19.0	30.6	23.9	13.8	(Miller and Charles, 1991)
AN-SYN	10675	-	-	0.95	-	-	0.68	(Mikolov et al., 2013a)
AN-SEM	8869	-	-	1.03	-	-	0.74	-Same-As-Above-
TOEFL	80	-	-	8.13	-	-	6.63	(Landauer and Dumais, 1997)

Table 2: List of test datasets used. The columns headed $\sigma^r_{p_0}$ contain *MRDS* values. The rows for accuracy based test sets contain σ_{p_0} which does not depend on r. See § 4.1 for details.

items, produced respectively by algorithms A and B, and then a list of gold ratings T. Let r_{AT}, r_{BT} and r_{AB} denote the Spearman correlations between $A:T$, $B:T$ and $A:B$ respectively. Let $\hat{r}_{AT}, \hat{r}_{BT}, \hat{r}_{AB}$ be their empirical estimates and assume that $\hat{r}_{BT} > \hat{r}_{AT}$ without loss of generality.

For word similarity datasets we define $\sigma^r_{p_0}$ as the MRDS, such that it satisfies the following proposition:

$$(r_{AB} < r) \wedge (|\hat{r}_{BT} - \hat{r}_{AT}| < \sigma^r_{p_0}) \Longrightarrow \textit{pval} > p_0$$

. Here *pval* is the probability of the test statistic under the null hypothesis that $r_{AT} = r_{BT}$ found using the Steiger's test (Steiger, 1980). The above constraint ensures that as long as the correlation between the competing methods is less than r and the difference between the correlations of the scores of the competing methods to the gold ratings is less than $\sigma^r_{p_0}$, then the pvalue of the null hypothesis will be greater than p_0. We can then ask what we consider a reasonable upper bound on the agreement of ratings produced by competing algorithms: for instance two algorithms correlating above 0.9 might not be considered meaningfully different. That leaves us with the second part of the predicate which ensures that as long as the difference between the correlations of the competing algorithms to the gold scores is less than $\sigma^r_{p_0}$ then the null hypothesis is more likely than p_0.

We can find $\sigma^r_{p_0}$ as follows: Let *stest* denote Steiger's test predicate which satisfies the following:

$$\textit{stest-p}(\hat{r}_{AT}, \hat{r}_{BT}, r_{AB}, p_0, n) \Longrightarrow \textit{pval} < p_0$$

Once we define this predicate then we can use it to set up an optimistic problem where our aim is to find $\sigma^r_{p_0}$ by solving the following:

$$\sigma^r_{p_0} = \min\{\sigma | \forall\, 0 < r' < 1\ \textit{stest-p}(r', \min(r'+\sigma, 1), r, p_0, n)\}$$

Note that MRDS is a liberal threshold and it only guarantees that differences in correlations below that threshold can never be statistically significant (under the given parameter settings). MRDS might optimistically consider some differences as significant when they are not, but it is at least useful in reducing some of the noise in the evaluations. The values of $\sigma^r_{p_0}$ are shown in Table 2.

For the accuracy based test-sets we found MRDS$=\sigma_{p_0}$ that satisfied the following:

$$0 < (\hat{\theta}_B - \hat{\theta}_A) < \sigma_{p_0} \Longrightarrow \mathrm{p}(\theta_B \leq \theta_A) > p_0$$

Specifically, we calculated the posterior probability $\mathrm{p}(\theta_B \leq \theta_A)$ with a flat prior of $\beta(1,1)$ to solve the following:[4] $\sigma_{p_0} = \min\{\sigma | \forall\, 0 < \theta < \min(1-\sigma, 0.9)$ $\mathrm{p}(\theta_B \leq \theta_A | \hat{\theta}_A = \theta, \hat{\theta}_B = \theta + \sigma, n) < p_0\}$ Here θ_A and θ_B

[4] This instead of using McNemar's test (McNemar, 1947) since the Bayesian approach is tractable and more direct. A calculation with $\beta(0.5, 0.5)$ as the prior changed $\sigma_{0.5}$ from 6.63 to 6.38 for the TOEFL dataset but did not affect MRDS for the AN-SEM and AN-SYN datasets.

are probability of correctness of algorithms A, B and $\hat{\theta}_A$, $\hat{\theta}_B$ are observed empirical accuracies.

Unfortunately there are no widely reported train-test splits of the above datasets, leading to potential concerns of *soft supervision* (hyper-parameter tuning) on these evaluations, both in our own work and throughout the existing literature. We report on the resulting impact of various parameterizations, and our final results are based on a single set of parameters used across all evaluation sets.

5 Experiments and Results

We wanted to answer the following questions through our experiments: (1) How do hyper-parameters affect performance? (2) What is the contribution of the multiple sources of data to performance? (3) How does the performance of MVLSA compare with other methods? For brevity we show tuning runs only on the larger datasets. We also highlight the top performing configurations in bold using the small threshold values in column $\sigma_{0.05}^{0.09}$ of Table 2.

Effect of Hyper-parameters f_j: We modeled the preprocessing function f_j as the composition of two functions, $f_j = n_j \circ t_j$. n_j represents nonlinear preprocessing that is usually employed with LSA. We experimented by setting n_j to be: identity; logarithm of count plus one; and the fourth root of the count. t_j represents the truncation of columns and can be interpreted as a type of regularization of the raw counts themselves through which we prune away the noisy contexts. Decrease in t_j also reduces the influence of views that have a large number of context columns and emphasizes the sparser views. Table 3 and Table 4 show the results.

Test Set	Log	Count	Count$^{\frac{1}{4}}$
MEN	67.5	59.7	**70.7**
RW	31.1	25.3	**37.8**
SCWS	64.2	58.2	**66.6**
AN-SYN	45.7	21.1	**53.6**
AN-SEM	25.4	15.9	**38.7**

Table 3: Performance versus n_j, the non linear processing of cooccurrence counts. $t = 200K,\ m = 500,\ v = 16,\ k = 300$. All the top configurations determined by $\sigma_{0.05}^{0.09}$ are in bold font.

Test Set	6.25K	12.5K	25K	50K	100K	200K
MEN	70.2	**71.2**	**71.5**	**71.6**	**71.2**	**70.7**
RW	**41.8**	**41.7**	**41.5**	**40.9**	39.6	37.8
SCWS	**67.1**	**67.3**	**67.1**	**67.0**	**66.9**	**66.6**
AN-SYN	59.2	**60.0**	**59.5**	58.4	56.1	53.6
AN-SEM	37.7	**38.6**	**39.4**	**39.2**	38.4	**38.7**

Table 4: Performance versus the truncation threshold, t, of raw cooccurrence counts. We used $n_j = \text{Count}^{\frac{1}{4}}$ and other settings were the same as Table 3.

m: The number of left singular vectors extracted after SVD of the preprocessed cooccurrence matrices can again be interpreted as a type of regularization, since the result of this truncation is that we find cooccurrence patterns only between the top left singular vectors. We set $m_j = max(d_j, m)$ with $m = [100, 300, 500]$. See table 5.

Test Set	100	200	300	500
MEN	65.6	68.5	**70.1**	**71.1**
RW	34.6	**36.0**	**37.2**	**37.1**
SCWS	64.2	**65.4**	**66.4**	**66.5**
AN-SYN	50.5	**56.2**	**56.4**	**56.4**
AN-SEM	24.3	31.4	34.3	**40.6**

Table 5: Performance versus m, the number of left singular vectors extracted from raw cooccurrence counts. We set $n_j = \text{Count}^{\frac{1}{4}},\ t = 100K,\ v = 25,\ k = 300$.

k: Table 6 demonstrates the variation in performance versus the dimensionality of the learnt vector representations of the words. Since the dimensions of the MVLSA representations are orthogonal to each other therefore creating lower dimensional representations is a trivial matrix slicing operation and does not require retraining.

Test Set	10	50	100	200	300	500
MEN	49.0	67.0	**69.7**	**70.2**	**70.1**	**69.8**
RW	28.8	33.3	35.0	35.2	**37.2**	**38.3**
SCWS	57.8	64.4	**65.2**	**66.1**	**66.4**	**65.1**
AN-SYN	9.0	41.2	52.2	55.4	**56.4**	54.4
AN-SEM	2.5	21.8	34.8	**35.8**	34.3	33.8

Table 6: Performance versus k, the final dimensionality of the embeddings. We set $m = 300$ and other settings were same as Table 5.

v: Expression 12 describes a method to set W_j. We experimented with a different, more global,

heuristic to set $[W_j]_{ii} = (K_{ww} \geq v)$, essentially removing all words that did not appear in v views before doing GCCA. Table 7 shows that changes in v are largely inconsequential for performance.

Test Set	16	17	21	25	29
MEN	**70.4**	**70.4**	**70.2**	**70.1**	**70.0**
RW	**39.9**	**38.8**	**39.7**	37.2	33.5
SCWS	**67.0**	**66.8**	**66.5**	**66.4**	**65.7**
AN-SYN	**56.0**	**55.8**	**55.9**	**56.4**	**56.0**
AN-SEM	**34.6**	**34.3**	**34.0**	**34.3**	**34.3**

Table 7: Performance versus minimum view support threshold v, The other hyperparameters were $n_j = \text{Count}^{\frac{1}{4}}$, $m = 300$, $t = 100K$. Though a clear best setting did not emerge, we chose $v = 25$ as the middle ground.

r_j: The regularization parameter ensures that all the inverses exist at all points in our method. We found that the performance of our procedure was invariant to r over a large range from 1 to 1e-10. This was because even the 1000th singular value of our data was much higher than 1.

Contribution of different sources of data Table 8 shows an ablative analysis of performance where we remove individual views or some combination of them and measure the performance. It is clear by comparing the last column to the second column that adding in more views improves performance. Also we can see that the Dependency based views and the Bitext based views give a larger boost than the morphology and FrameNet based views, probably because the latter are so sparse.

Comparison to other word representation creation methods There are a large number of methods of creating representations both multilingual and monolingual. There are many new methods such as by Yu and Dredze (2014), Faruqui et al. (2014), Hill and Korhonen (2014), and Weston et al. (2014) that are performing multiview learning and could be considered here as baselines: however it is not straightforward to use those systems to handle the variety of data that we are using. Therefore, we directly compare our method to the Glove and the SkipGram model of Word2Vec as the performance of those systems is considered state of the art. We trained these two systems on the English portion of the *Polyglot* Wikipedia dataset.[5] We also combined their outputs using MVLSA to create *MV-G-WSG)* embeddings.

We trained our best MVLSA system with data from all views and by using the individual best settings of the hyper-parameters. Specifically the configuration we used was as follows: $n_j = \text{Count}^{\frac{1}{4}}, t = 12.5K, m = 500, k = 300, v = 16$. To make a fair comparison we also provide results where we used only the views derived from the *Polyglot* Wikipedia corpus. See column *MVLSA (All Views)* and *MVLSA (Wiki)* respectively. It is clearly visible that MVLSA on the monolingual data itself is competitive with Glove but worse than Word2Vec on the word similarity datasets and it is substantially worse than both the systems on the AN-SYN and AN-SEM datasets. However with the addition of multiple views MVLSA makes substantial gains, shown in column *MV Gain*, and after consuming the Glove and WSG embeddings it again improves performance by some margins, as shown in column *G-WSG Gain*, and outperforms the original systems. Using GCCA itself for system combination provides closure for the MVLSA algorithm since multiple distinct approaches can now be simply fused using this method. Finally we contrast the Spearman correlations r_s with Glove and Word2Vec before and after including them in the GCCA procedure. The values demonstrate that including Glove and WSG during GCCA actually increased the correlation between them and the learnt embeddings, which supports our motivation for performing GCCA in the first place.

6 Previous Work

Vector space representations of words have been created using diverse frameworks including Spectral methods (Dhillon et al., 2011; Dhillon et al., 2012), [6] Neural Networks (Mikolov et al., 2013b; Collobert and Lebret, 2013), and Random Projections (Ravichandran et al., 2005; Bhagat and Ravichan-

[5] We explicitly provided the vocabulary file to Glove and Word2Vec and set the truncation threshold for Word2Vec to 10. Glove was trained for 25 iterations. Glove was provided a window of 15 previous words and Word2Vec used a symmetric window of 10 words.

[6] `cis.upenn.edu/~ungar/eigenwords`

Test Set	All Views	!Framenet	!Morphology	!Bitext	!Wikipedia	!Dependency	!Morphology !Framenet	!Morphology !Framenet !Bitext
MEN	**70.1**	**69.8**	**70.1**	**69.9**	46.4	68.4	**69.5**	68.4
RW	**37.2**	**36.4**	**36.1**	32.2	11.6	34.9	34.1	27.1
SCWS	**66.4**	**65.8**	**66.3**	64.2	54.5	**65.5**	**65.2**	60.8
AN-SYN	**56.4**	**56.3**	**56.2**	51.2	37.6	50.5	54.4	46.0
AN-SEM	34.3	34.3	34.3	**36.2**	4.1	35.3	34.5	30.6

Table 8: Performance versus views removed from the multiview GCCA procedure. !Framenet means that the view containing counts derived from Frame semantic dataset was removed. Other columns are named similarly. The other hyperparameters were $n_j = \text{Count}^{\frac{1}{4}},\ m = 300,\ t = 100K,\ v = 25,\ k = 300$.

Test Set	Glove	WSG	MV G-WSG	MVLSA Wiki	MVLSA All Views	MVLSA Combined	MV Gain	G-WSG Gain	r_s MVLSA Glove	r_s MVLSA WSG	r_s MV-G-WSG Glove	r_s MV-G-WSG WSG
MEN	70.4	73.9	**76.0**	71.4	71.2	**75.8**	-0.2	$4.6^{\dagger}$	71.9	89.1	85.8	92.3
RW	28.1	32.9	37.2	29.0	**41.7**	**40.5**	$12.7^{\dagger}$	-1.2	72.3	74.2	80.2	75.6
SCWS	54.1	65.6	60.7	61.8	**67.3**	**66.4**	$5.5^{\dagger}$	-0.9	87.1	94.5	91.3	96.3
SIMLEX	33.7	36.7	41.1	34.5	**42.4**	**43.9**	$7.9^{\dagger}$	1.5	62.4	78.2	79.3	86.0
WS	58.6	**70.8**	**67.4**	**68.0**	**70.8**	**70.1**	$2.8^{\dagger}$	-0.7	72.3	88.1	81.8	91.8
MTURK	**61.7**	**65.1**	59.8	59.1	59.7	**62.9**	0.6	3.2	80.0	87.7	87.3	92.5
WS-REL	53.4	**63.6**	59.6	60.1	**65.1**	**63.5**	$5.0^{\dagger}$	-1.6	58.2	81.0	69.6	85.3
WS-SEM	69.0	**78.4**	**76.1**	**76.8**	**78.8**	**79.2**	2.0	0.4	74.4	90.6	83.9	94.0
RG	**73.8**	**78.2**	**80.4**	71.2	**74.4**	**80.8**	3.2	$6.4^{\dagger}$	80.3	90.6	91.8	92.9
MC	**70.5**	**78.5**	**82.7**	**76.6**	**75.9**	**77.7**	-0.7	2.8	80.1	94.1	91.4	95.8
AN-SYN	61.8	59.8	51.0	42.7	60.0	**64.3**	$17.3^{\dagger}$	$4.3^{\dagger}$				
AN-SEM	**80.9**	73.7	73.5	36.2	38.6	77.2	$2.4^{\dagger}$	$38.6^{\dagger}$				
TOEFL	**83.8**	81.2	**86.2**	78.8	**87.5**	**88.8**	$8.7^{\dagger}$	1.3				

Table 9: Comparison of Multiview LSA against Glove and WSG(Word2Vec Skip Gram). Using $\sigma_{0.05}^{0.9}$ as the threshold we highlighted the top performing systems in bold font. † marks significant increments in performance due to use of multiple views in the *Gain* columns. The r_s columns demonstrate that GCCA increased pearson correlation.

dran, 2008; Chan et al., 2011). [7] They have been trained using either one (Pennington et al., 2014) [8] or two sources of cooccurrence statistics (Zou et al., 2013; Faruqui and Dyer, 2014; Bansal et al., 2014; Levy and Goldberg, 2014) [9] or using multi-modal data (Hill and Korhonen, 2014; Bruni et al., 2012).

Dhillon et al. (2011) and Dhillon et al. (2012) were the first to use CCA as the primary method to learn vector representations and Faruqui and Dyer (2014) further demonstrated that incorporating bilingual data through CCA improved performance. More recently this same phenomenon was reported by Hill et al. (2014a) through their experiments over neural representations learnt from MT systems. Various other researchers have tried to improve the performance of their paraphrase systems or vector space models by using diverse sources of information such as bilingual corpora (Bannard and Callison-Burch, 2005; Huang et al., 2012; Zou et al., 2013),[10] structured datasets (Yu and Dredze, 2014; Faruqui et al., 2014) or even tagged images (Bruni

[7] code.google.com/p/word2vec,metaoptimize.com/projects/wordreprs

[8] nlp.stanford.edu/projects/glove

[9] ttic.uchicago.edu/~mbansal/data/syntacticEmbeddings.zip, cs.cmu.edu/~mfaruqui/soft.html

[10] An example of complementary views: Chan et al. (2011) observed that monolingual distributional statistics are susceptible to conflating antonyms, where bilingual data is not; on the other hand bilingual statistics are susceptible to noisy alignments, where monolingual data is not.

et al., 2012). However, most previous work[11] did not adopt the general, simplifying view that all of these sources of data are just cooccurrence statistics coming from different sources with underlying latent factors.[12]

Bach and Jordan (2005) presented a probabilistic interpretation for CCA. Though they did not generalize it to include GCCA we believe that one could give a probabilistic interpretation of *MAX-VAR GCCA*. Such a probabilistic interpretation would allow for an online-generative model of lexical representations, which unlike methods like Glove or LSA would allows us to naturally perplexity or generate sequences. We also note that Vía et al. (2007) presented a neural network model of GCCA and adaptive/incremental GCCA. To the best of our knowledge both of these approaches have not been used for word representation learning.

CCA is also an algorithm for multi-view learning (Kakade and Foster, 2007; Ganchev et al., 2008) and when we view our work as an application of multi-view learning to NLP, this follows a long chain of effort started by Yarowsky (1995) and continued with *Co-Training* (Blum and Mitchell, 1998), *CoBoosting* (Collins and Singer, 1999) and *2 view perceptrons* (Brefeld et al., 2006).

7 Conclusion and Future Work

While previous efforts demonstrated that incorporating two views is beneficial in word-representation learning, we extended that thread of work to a logical extreme and created *MVLSA* to learn distributed representations using data from 46 views![13] Through evaluation of our induced representations, shown in Table 9, we demonstrated that the MVLSA algorithm is able to leverage the information present in multiple data sources to improve performance on a battery of tests against state of the art baselines. In order to perform MVLSA on large vocabularies with up to 500K words we presented a fast scalable algorithm. We also showed that a close variant of the Glove objective proposed by Pennington et al. (2014) could be derived as a heuristic for handling missing data under the MVLSA framework. In order to better understand the benefit of using multiple sources of data we performed MVLSA using views derived only from the monolingual Wikipedia dataset thereby providing a more principled alternative of LSA that removes the need for heuristically combining word-word cooccurrence matrices into a single matrix. Finally, while surveying the literature we noticed that not enough emphasis was being given towards establishing the significance of comparative results and proposed a method, *(MRDS)*, to filter out insignificant comparative gains between competing algorithms.

Future Work Column *MVLSA Wiki* of Table 9 shows us that MVLSA applied to monolingual data has mediocre performance compared to the baselines of Glove and Word2Vec on word similarity tasks and performs surprisingly worse on the ANSEM dataset. We believe that the results could be improved by (1) either using recent methods for handling missing values mentioned in footnote 1 or by using the heuristic count dependent non-linear weighting mentioned by Pennington et al. (2014) and that sits well within our framework as exemplified in Expression 12 (2) by using even more views, which look at the future words as well as views that contain PMI values. Finally, we note that Table 8 shows that certain datasets can actually degrade performance over certain metrics. Therefore we are exploring methods for performing discriminative optimization of weights assigned to views, for purposes of task-based customization of learned representations.

Acknowledgments

This material is based on research sponsored by the Defense Advanced Research Projects Agency (DARPA) under the Deep Exploration and Filtering of Text (DEFT) Program, agreement number FA8750-13-2-001, as well as the National Science Foundation (NSF), agreement number BCS-1344269. We also thank Juri Ganitkevitch for providing the word aligned bitext corpus.

[11] Ganitkevitch et al. (2013) did employ a rich set of diverse cooccurrence statistics in constructing the initial PPDB, but without a notion of "training" a joint representation beyond random projection to a binary vector subspace (bit-signatures).

[12] Note that while Faruqui et al. (2014) performed belief propagation over a graph representation of their data, such an undirected weighted graph can be viewed as an adjacency matrix, which is then also a cooccurrence matrix.

[13] Code and data available at `www.cs.jhu.edu/~prastog3/mvlsa`

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of NAACL-HLT*. ACL.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of CoNLL*. ACL.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. 2014. Tensor decompositions for learning latent variable models. *JMLR*, 15.

Raman Arora and Karen Livescu. 2012. Kernel CCA for multi-view learning of acoustic features using articulatory measurements. *MLSLP*.

Francis R Bach and Michael I Jordan. 2005. A probabilistic interpretation of canonical correlation analysis. *Technical Report 688, Department of Statistics, University of California, Berkeley*.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of ACL*. ACL.

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *Proceedings of ACL*. ACL.

Rahul Bhagat and Deepak Ravichandran. 2008. Large scale acquisition of paraphrases for learning surface patterns. In *Proceedings of ACL-HLT*.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of COLT*. ACM.

Matthew Brand. 2002. Incremental singular value decomposition of uncertain data with missing values. In *Computer Vision—ECCV 2002*, pages 707–720. Springer.

Matthew Brand. 2006. Fast low-rank modifications of the thin singular value decomposition. *Linear algebra and its applications*, 415(1).

Ulf Brefeld, Thomas Gärtner, Tobias Scheffer, and Stefan Wrobel. 2006. Efficient co-regularised least squares regression. In *Proceedings of ICML*. ACM.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of ACL*. ACL.

J Douglas Carroll. 1968. Generalization of canonical correlation analysis to three or more sets of variables. In *Proceedings of APA*, volume 3.

Tsz Ping Chan, Chris Callison-Burch, and Benjamin Van Durme. 2011. Reranking bilingually extracted paraphrases using monolingual distributional similarity. In *Proceedings of EMNLP Workshop: GEMS*.

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *Proceedings of EMNLP*. ACL.

Ronan Collobert and Rémi Lebret. 2013. Word embeddings through hellinger pca. Technical report, Idiap.

Paramveer Dhillon, Dean Foster, and Lyle Ungar. 2011. Multi-view learning of word embeddings via CCA. In *Procesdings of NIPS*.

Paramveer Dhillon, Jordan Rodu, Dean P Foster, and Lyle H Ungar. 2012. Two step CCA: A new spectral method for estimating vector models of words. In *Proceedings of ICML*. ACM.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of EACL*.

Manaal Faruqui, Jesse Dodge, Sujay Jauhar, Chris Dyer, Eduard Hovy, and Noah Smith. 2014. Retrofitting word vectors to semantic lexicons. In *Proceedings of the deep learning and representation learning workshop, NIPS*.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of WWW*. ACM.

Kuzman Ganchev, Joao Graca, John Blitzer, and Ben Taskar. 2008. Multi-view learning over structured and non-identical outputs. In *Proceedings of UAI*.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *Proceedings of NAACL-HLT*.

Nizar Habash and Bonnie Dorr. 2003. Catvar: A database of categorial variations for english. In *Proceedings of MT Summit*.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements Of Statistical Learning*, volume 2. Springer.

Felix Hill and Anna Korhonen. 2014. Learning abstract concept embeddings from multi-modal data: Since you probably can't see what i mean. In *Proceedings of EMNLP*. ACL.

Felix Hill, KyungHyun Cho, Sebastien Jean, Coline Devin, and Yoshua Bengio. 2014a. Not all neural embeddings are born equal. *arXiv preprint arXiv:1410.0718*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014b. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *arXiv preprint arXiv:1408.3456*.

Paul Horst. 1961. Generalized canonical correlations and their applications to experimental data. *Journal of Clinical Psychology*, 17(4).

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of ACL*. ACL.

Sham M Kakade and Dean P Foster. 2007. Multi-view regression via canonical correlation analysis. In *Learning Theory*. Springer.

Jon R Kettenring. 1971. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of ACL*. ACL.

Minh-Thang Luong, Richard Socher, and Christopher D. Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of CoNLL*. ACL.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2).

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT*, pages 746–751.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1).

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of english. *Natural Language Engineering*, 7(03).

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of NAACL Workshop: AKBC-WEKEX*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: global vectors for word representation. In *Proceedings of EMNLP*. ACL.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: Computing word relatedness using temporal semantic analysis. In *Proceedings of WWW*. ACM.

Pushpendre Rastogi and Benjamin Van Durme. 2014. Augmenting framenet via PPDB. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*. ACL.

Deepak Ravichandran, Patrick Pantel, and Eduard Hovy. 2005. Randomized algorithms and nlp: Using locality sensitive hash functions for high speed noun clustering. In *Proceedings of ACL*.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10).

Dmitry Savostyanov. 2014. Efficient way to find svd of sum of projection matrices? MathOverflow. URL:http://mathoverflow.net/q/178573 (version: 2014-08-14).

Karthik Sridharan and Sham M Kakade. 2008. An information theoretic framework for multi-view learning. In *Proceedings of COLT*.

James H Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2).

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of AI Research*, 37(1).

Michel Van De Velden and Tammo HA Bijmolt. 2006. Generalized canonical correlation analysis of matrices with missing rows: a simulation study. *Psychometrika*, 71(2).

Michel van de Velden and Yoshio Takane. 2012. Generalized canonical correlation analysis with missing values. *Computational Statistics*, 27(3).

Javier Vía, Ignacio Santamaría, and Jesús Pérez. 2007. A learning algorithm for adaptive canonical correlation analysis of several data sets. *Neural Networks*, 20(1).

Jason Weston, Sumit Chopra, and Keith Adams. 2014. #tagspace: Semantic embeddings from hashtags. In *Proceedings of EMNLP*, Doha, Qatar. ACL.

Terry Winograd. 1972. Understanding natural language. *Cognitive psychology*, 3(1):1–191.

David Yarowsky. 1995. Unsupervised WSD rivaling supervised methods. In *Proceedings of ACL*. ACL.

Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *Proceedings of ACL*. ACL.

Will Zou, Richard Socher, Daniel Cer, and Christopher Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of EMNLP*. ACL.