Multiview LSA: Representation Learning Via Generalized CCA

Pushpendre Rastogi[1], Benjamin Van Durme[1,2], Raman Arora[1]

[1]Center for Language and Speech Processing, JHU
[2]Human Language Technology Center of Excellence

The Center For Language and Speech Processing at the Johns Hopkins University

Johns Hopkins University

Multiview LSA

- Represent datasets (linguistic or otherwise) as matrices, such that **each matrix is a view** of a word/phrase.
- Use **Max-Var GCCA** to create embeddings.
- Use **incremental SVD** so that the method can scale to handle millions of words/phrases and hundreds of views, where a view can be either a sparse or a dense matrix.
- **Handle missing values** instead of ignoring

Max-Var GCCA

LSA is an application of **PCA** to a *single* term-document cooccurrence matrix. **CCA** learns linear projections that are maximally correlated to each other from *two views*, **Generalized CCA** is a family of extensions of CCA to maximize correlation across *multiple views*.

One variant of GCCA called **MAX-VAR GCCA** induces an auxilliary representation G that is maximally correlated to linear projections of the views in terms of sum of squared correlations [1, 2].

$$G = \text{eig}\left(\sum_{j=1}^{J} P_j\right)$$
$$\text{Where, } P_j = X_j(X_j^\top X_j)^{-1}X_j^\top$$

Handling Missing Values

Sparse cooccurrence matrices contain plenty of missing values that cripple the performance of methods that rely on spectral decompositions. We address this sparsity by optimizing our representations only on the observed rows using **a variant of MAX-Var GCCA** presented by [3].

$$G = \text{eig}\left((\sum_j K_j)^{-\frac{1}{2}}(\sum_{j=1}^{J} P_j)(\sum_j K_j)^{-\frac{1}{2}}\right) \quad (1)$$

where $[K_j]_{ii} = 1$ if row i of view j is observed and zero otherwise.

Further Information

- Visit: www.cs.jhu.edu/~prastog3/mvlsa
- Email: pushpendre@jhu.edu

Abstract

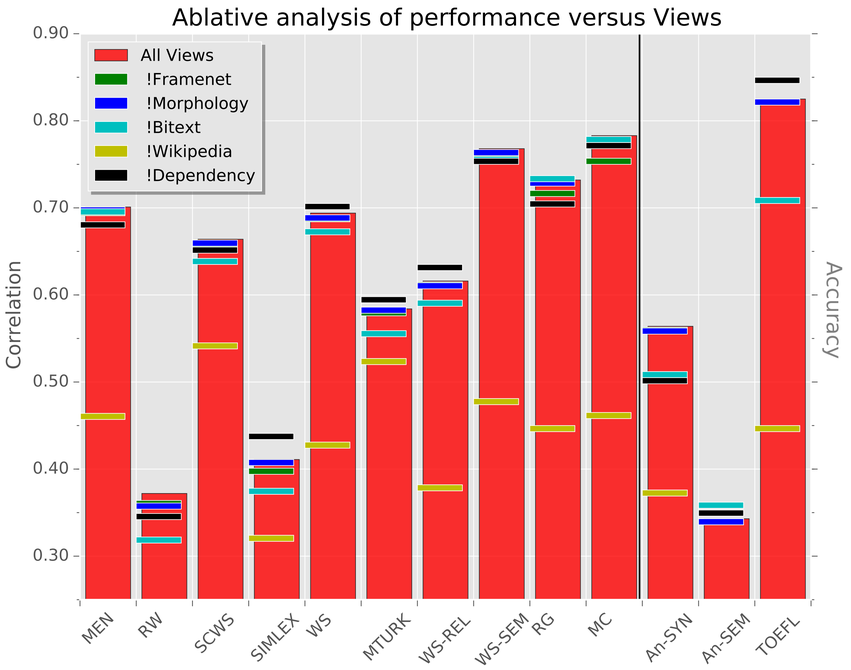
Multiview LSA is a way of utilizing **hundreds of data sources** to learn representations for **millions of words/phrases** that outperform baselines like **Word2Vec and Glove** [4, 5].

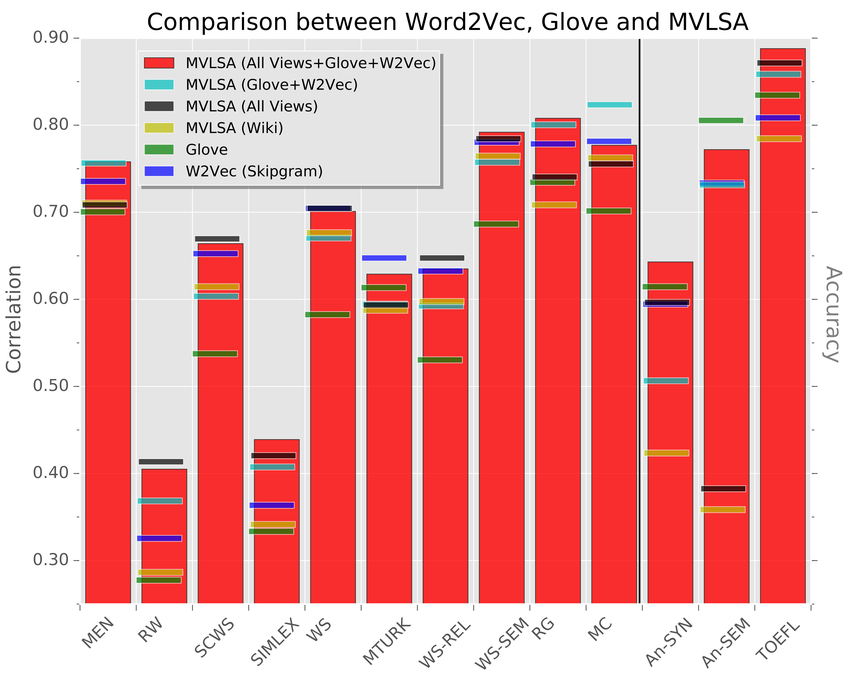
Training Datasets

- 15 word history from Polyglot English Wikipedia Corpus
- Word alignment statistics from 6 Word Aligned bitext corpora (Arabic, Czech, German, Spanish, French, Chinese)
- Parent child cooccurrence events for 22 dependency relations from Annotated GigaWord
- Framenet Lexical Units augmented with PPDB paraphrases
- Morphological information from Catvar, Morpha, Morphg and Morphy
- Embeddings generated by Glove and Word2Vec

Test Set	Size	$\sigma^{0.9}_{0.05}$
MEN	3000	1.3
RW	2034	1.6
SCWS	2003	1.6
SIMLEX	999	2.3
WS	353	3.9
MTURK	287	4.3
WS-REL	252	4.6
WS-SEM	203	5.1
RG	65	9.2
MC	30	13.8
T-SYN	10675	0.68
T-SEM	8869	0.74
TOEFL	80	6.63

Table 1: Common test sets and associated MRDS values. MRDS= $\sigma^{0.9}_{0.05}$ measures the minimum required difference between two algorithms for that difference to be significant with a pval of 0.05 assuming that the maximum correlation between the ratings produced by the competing algorithms is 0.9.





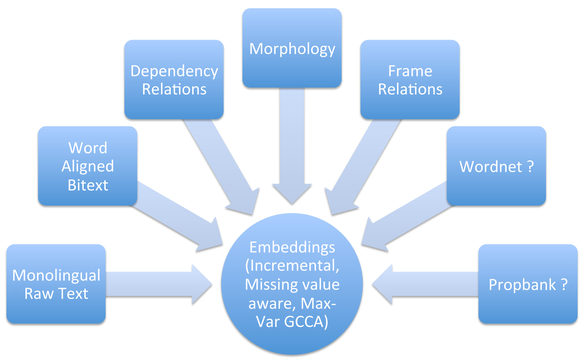
Unifying Prior Work

We could approximately mimic the objective of [5] by changing the reweighting terms in our method for handling missing values as follows:

$$\underset{G,U_j}{\text{minimize:}} \quad \sum_{j=1}^{J} \|W_j K_j (G - X_j U_j)\|_F^2 \quad (2)$$
$$\text{subject to: } G^\top G = I$$

where

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



References

[1] Douglas Carroll. Generalization of canonical correlation analysis to three or more sets of variables. *APA*, 1968.

[2] Jon Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 1971.

[3] Michel Van De Velden and Tammo Bijmolt. Generalized canonical correlation analysis of matrices with missing rows: a simulation study. *Psychometrika*, 2006.

[4] Tomas Mikolov et. al. Distributed representations of words and phrases and their compositionality. *NIPS*, 2013.

[5] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: global vectors for word representation. *EMNLP*. ACL, 2014.

Acknowledgements

This material is based on research sponsored by Defense Advanced Research Projects Agency (DARPA) under the Deep Exploration and Filtering of Text (DEFT) Program (Agreement number FA8750-13-2-0017).