Supplementary Material:
Multi-View Learning of Word Embeddings via CCA

Our goal is to find a $v \times k$ matrix $\mathbf{A}$ that maps each of the v words in the vocabulary to a k-dimensional state vector. We will show that the $\mathbf{A}$ we find preserves the information in our data and allows a significant data reduction.

Let $\mathbf{L}$ be an $n \times hv$ matrix giving the words in the left context of each of the n tokens, where the context is of length h, $\mathbf{R}$ be the corresponding $n \times hv$ matrix for the right context, and $\mathbf{W}$ be an $n \times v$ matrix of indicator functions for the words themselves.

We will use three assumptions at various points in our proof:

Assumption 1. $\mathbf{L}$*,* $\mathbf{W}$*, and* $\mathbf{R}$ *come from a rank* k *HMM i.e it has a rank* k *observation matrix and a rank* k *transition matrix both of which have the same domain.*

For example, if the dimension of the hidden state is k and the vocabulary size is v then the observation matrix, which is $k \times v$, has rank k. This rank condition is similar to the one used by Siddiqi et al. (2010).

Assumption 1A. *For the three views,* $\mathbf{L}$*,* $\mathbf{W}$ *and* $\mathbf{R}$ *assume that there exists a "hidden state H" of dimension* $n \times k$*, where each row* H_i *has the same non-singular variance-covariance matrix and such that* $\mathbb{E}(L_i|H_i) = H_i\boldsymbol{\beta}_L^T$ *and* $\mathbb{E}(R_i|H_i) = H_i\boldsymbol{\beta}_R^T$ *and* $\mathbb{E}(W_i|H_i) = H_i\boldsymbol{\beta}_W^T$ *where all* $\boldsymbol{\beta}$*'s are of rank* k*, where* L_i*,* R_i *and* W_i *are the rows of* $\mathbf{L}$*,* $\mathbf{R}$ *and* $\mathbf{W}$ *respectively.*

This assumption actually follows from the previous one.

Assumption 2. $\rho(\mathbf{L}, \mathbf{W})$*,* $\rho(\mathbf{L}, \mathbf{R})$ *and* $\rho(\mathbf{W}, \mathbf{R})$ *all have rank* k*, where* $\rho(\mathbf{X_1}, \mathbf{X_2})$ *is the expected correlation between* $\mathbf{X_1}$ *and* $\mathbf{X_2}$*.*

This is a rank condition similar to that in Hsu et al. (2009).

Assumption 3. $\rho([\mathbf{L}, \mathbf{R}], \mathbf{W})$ *has* k *distinct singular values.*

This assumption just makes the proof a little cleaner, since if there are repeated singular values, then the singular vectors are not unique. Without it, we would have to phrase results in terms of subspaces with identical singular values.

We also need to define the CCA function that computes the left and right singular vectors for a pair of matrices:

Definition 1 (CCA)**.** *Compute the CCA between two matrices* $\mathbf{X_1}$ *and* $\mathbf{X_2}$*. Let* $\mathbf{\Phi_{X_1}}$ *be a matrix containing the* d *largest singular vectors for* $\mathbf{X_1}$ *(sorted from the largest on down). Likewise for* $\mathbf{\Phi_{X_2}}$*. Define the function* $CCA_d(\mathbf{X_1}, \mathbf{X_2}) = [\mathbf{\Phi_{X_1}}, \mathbf{\Phi_{X_2}}]$*. When we want just one of these* $\mathbf{\Phi}$*'s, we will use* $CCA_d(\mathbf{X_1}, \mathbf{X_2})_{left} = \mathbf{\Phi_{X_1}}$ *for the left singular vectors and* $CCA_d(\mathbf{X_1}, \mathbf{X_2})_{right} = \mathbf{\Phi_{X_2}}$ *for the right singular vectors.*

Note that the resulting singular vectors, $[\Phi_{X_1}, \Phi_{X_2}]$ can be used to give two redundant estimates, $X_1\Phi_{X_1}$ and $X_2\Phi_{X_2}$ of the "hidden" state relating X_1 and X_2, if such a hidden state exists.

Definition 2. *Define the symbol "*$\approx$*" to mean*

$$\mathbf{X_1} \approx \mathbf{X_2} \iff \lim_{n\to\infty} \mathbf{X_1} = \lim_{n\to\infty} \mathbf{X_2}$$

where n *is the sample size.*

Lemma 1. *Define A by the following limit of the right singular vectors:*

$$CCA_k([\mathbf{L}, \mathbf{R}], \mathbf{W})_{\mathit{right}} \approx \mathbf{A}.$$

Under assumptions 2, 3 and 1A, such that if $CCA_k(\mathbf{L}, \mathbf{R}) \equiv [\mathbf{\Phi}_L, \mathbf{\Phi}_R]$ then we have

$$CCA_k([\mathbf{L\Phi}_L, \mathbf{R\Phi}_R], \mathbf{W})_{\mathit{right}} \approx \mathbf{A}.$$

This lemma shows that instead of finding the CCA between the full context and the words, we can take the CCA between the Left and Right contexts, estimate a k dimensional state from them, and take the CCA of that state with the words and get the same result.

Proof:

By assumption 1A, we see that:

$$\mathbb{E}(\mathbf{L}\boldsymbol{\beta}_{\mathbf{L}}|\mathbf{H}) = \mathbf{H}\boldsymbol{\beta}_L^T\boldsymbol{\beta}_L$$

and

$$\mathbb{E}(\mathbf{R}\boldsymbol{\beta}_{\mathbf{R}}|\mathbf{H}) = \mathbf{H}\boldsymbol{\beta}_R^T\boldsymbol{\beta}_R$$

Since, again by assumption 1A both of the $\boldsymbol{\beta}$ matrixes have full rank, $\boldsymbol{\beta}_L^T\boldsymbol{\beta}_L$ is a $k \times k$ matrix of rank k, and likewise for $\boldsymbol{\beta}_R^T\boldsymbol{\beta}_R$. So

$$\mathbb{E}(\boldsymbol{\beta}_R^T\mathbf{R}^T\mathbf{L}\boldsymbol{\beta}_L|\mathbf{H}) = \boldsymbol{\beta}_R^T\boldsymbol{\beta}_R\mathbf{H}^T\mathbf{H}\boldsymbol{\beta}_L\boldsymbol{\beta}_L^T$$

So,

$$\boldsymbol{\beta}_R^T\mathbb{E}(\mathbf{R}^T\mathbf{L})\boldsymbol{\beta}_L = \boldsymbol{\beta}_R^T\boldsymbol{\beta}_R\mathbb{E}(\mathbf{H}^T\mathbf{H})\boldsymbol{\beta}_L\boldsymbol{\beta}_L^T$$

since $\boldsymbol{\beta}_R^T\boldsymbol{\beta}_R$, $\mathbb{E}(\mathbf{H}^T\mathbf{H})$ and $\boldsymbol{\beta}_L^T\boldsymbol{\beta}_L$ are all $k \times k$ full rank matrices, $\boldsymbol{\beta}_R$ and $\boldsymbol{\beta}_L$ span the same subspace as the singular values of the CCA between $\mathbf{L}$ and $\mathbf{R}$ since by assumption 2 they have rank k also. Similar arguments hold when relating $\mathbf{L}$ with $\mathbf{W}$ and when relating $\mathbf{R}$ with $\mathbf{W}$. Thus if $CCA_k(\mathbf{L}, \mathbf{R}), \mathbf{W}) \equiv [\Phi_L, \Phi_R]$,

$$CCA_k(\mathbf{L}\Phi_L, \mathbf{R}\Phi_R)_{\text{right}} \approx CCA_k([\mathbf{L}\boldsymbol{\beta}_L, \mathbf{R}\boldsymbol{\beta}_R], \mathbf{W})_{\text{right}}$$

(where we have used assumption 3 to ensure that not only are the subspaces the same, but that the actual singular vectors are the same.)

Finally by 3 we know that the rank of $CCA_k([\mathbf{L}, \mathbf{R}], \mathbf{W})_{\text{right}}$ is k we see that

$$CCA_k([\mathbf{L}\boldsymbol{\beta}_L, \mathbf{R}\boldsymbol{\beta}_R], \mathbf{W})_{\text{right}} \approx CCA_k([\mathbf{L}, \mathbf{R}], \mathbf{W})_{\text{right}}.$$

Calling this common limit $\mathbf{A}$ yields our result.
q.e.d.

Let $\tilde{\mathbf{A}}_h$ denote a matrix formed by stacking h copies of $\mathbf{A}$ on top of each other. Right multiplying $\mathbf{L}$ or $\mathbf{R}$ by $\tilde{\mathbf{A}}_h$ projects each of the words in that context into the k-dimensional reduced rank space.

The following theorem addresses the core of a new LR-MVL(II) algorithm, showing that there is an $\mathbf{A}$ which gives the desired dimensionality reduction.

Theorem 1. *Under assumptions 1, 2 and 3 there exists a unique matrix A such that if $CCA_k(\mathbf{L\tilde{A}_h}, \mathbf{R\tilde{A}_h}) \equiv [\tilde{\mathbf{\Phi}}_L, \tilde{\mathbf{\Phi}}_R]$ then*

$$CCA_k([\mathbf{L\tilde{A}_h\tilde{\Phi}_L}, \mathbf{R\tilde{A}_h\tilde{\Phi}_R}], \mathbf{W})_{\mathit{right}} \approx \mathbf{A}$$

where $\tilde{\mathbf{A}}_h$ is the stacked form of $\mathbf{A}$.

Proof: We start by noting that assumption 1 implies assumption 1A. Thus, the previous lemma follows. So we know

$$CCA_k([\mathbf{L}, \mathbf{R}], \mathbf{W})_{\text{right}} \approx CCA_k([\mathbf{L}\Phi_L, \mathbf{R}\Phi_R], \mathbf{W})_{\text{right}}$$

where, as usual, $CCA_k(\mathbf{L}, \mathbf{R}) \equiv [\Phi_L, \Phi_R]$, which allows us to define $\mathbf{A}$. This $\mathbf{A}$ has the property that the rank of $CCA(\mathbf{WA}, \mathbf{H})_{\text{left}}$ is the same as $CCA(\mathbf{W}, \mathbf{H})_{\text{left}}$ where $\mathbf{H}$ is the hidden state process associated with our data. Hence anything which is not in the domain of $\mathbf{A}$ won't have any correlation with $\mathbf{H}$ and hence no correlation with other observed states. So $\mathbf{L}$ and $\mathbf{L}\tilde{\mathbf{A}}_h$ have the same "information." More precisely,

$$[\tilde{\mathbf{A}}_h \tilde{\Phi_L}, \tilde{\mathbf{A}}_h \tilde{\Phi_R}] \approx CCA_k(\mathbf{L}, \mathbf{R})$$

where $CCA_k(\mathbf{L}\tilde{\mathbf{A}}_h, \mathbf{R}\tilde{\mathbf{A}}_h) \equiv [\tilde{\Phi_L}, \tilde{\Phi_R}]$ Putting this together with our first equation shows our desired result.
q.e.d.

References

Hsu, D, Kakade, S M., and Zhang, Tong. A spectral algorithm for learning hidden markov models. In *COLT*, 2009.

Siddiqi, S., Boots, B., and Gordon, G. J. Reduced-rank hidden Markov models. In *AISTATS-2010*, 2010.