THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Computer Science
College of Natural Sciences

Capturing Semantically Meaningful Word Dependencies with an Admixture of Poisson MRFs

David I. Inouye, Pradeep Ravikumar, Inderjit S. Dhillon

Admixture of Poisson MRFs (APM) [Inouye et al. 2014]

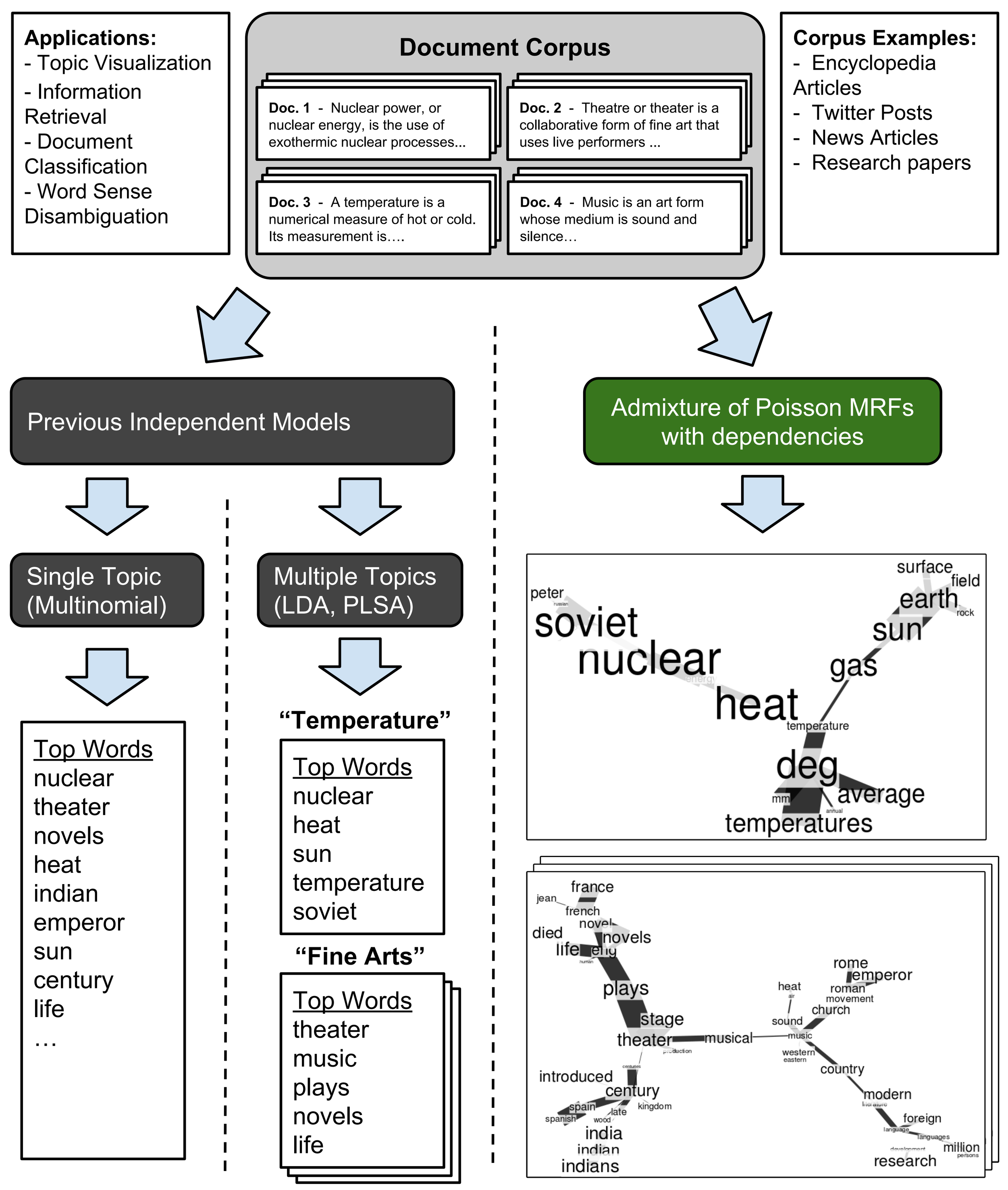


Figure: Previous topic models assume words are conditionally *independent* of each other given a topic, and thus, these previous models can only represent topics as a *list of words* ordered by frequency. However, an Admixture of Poisson MRFs (APM) can model dependencies between words and hence can represent each topic as a *graph over words*.

Open Problems in APM Model

1. High computational complexity of APM
   - No parallelism since optimizing jointly over all parameters
   - Slow convergence of proximal gradient descent
   - APM has $O(kp^2)$ parameters versus $O(kp)$ for LDA
2. Edge parameters of APM not directly evaluated
   - Previous metrics calculated word pair statistics for top words [Newman et al. 2010, Mimno et al. 2011, Aletras and Court 2013]
   - However, APM explicitly models dependencies between words
   - How can we semantically evaluate these dependencies?

Proposed Solutions

1. Parallel alternating Newton-like algorithm
   - Split into two convex problems
   - Demonstrate scaling at $p = 10,000$ and $n = 100,000$
   - Empirically, $O(knp^2)$ complexity to estimate $O(kn + kp^2)$ parameters
   - `http://bigdata.ices.utexas.edu/software/apm/`
2. Evocation metric that directly evaluates word pairs
   - Develop novel metric based on notion of *evocation* (which words "bring to mind" other words)

[Inouye et al. 2014] Inouye, D. I., Ravikumar, P., and Dhillon, I. S. Admixture of Poisson MRFs: A Topic Model with Word Dependencies. In *ICML*, 2014.

Background: Admixtures

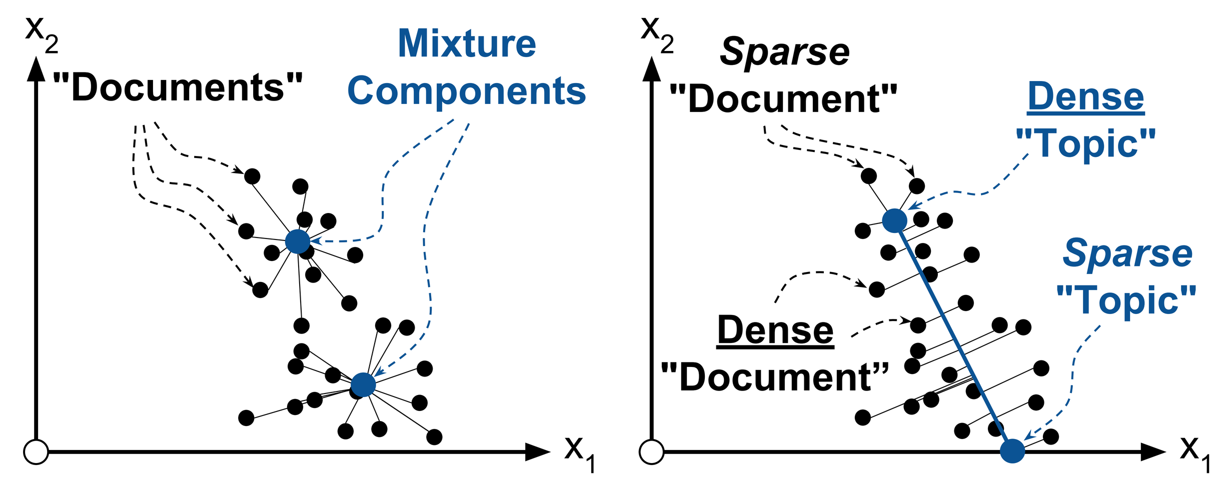


Figure: (Left) In *mixtures*, documents are drawn from exactly one component distribution. (Right) In *admixtures*, documents are drawn from a distribution whose parameters are a convex combination of component parameters.

The conditional distribution given the admixture weights and component distributions is merely the base distribution with parameters that are instance-specific mixtures of the component parameters:

$$\Pr_{\text{Admix.}}(\mathbf{x} \mid \mathbf{w}, \Phi) = \Pr_{\text{Base}}\left(\mathbf{x} \,\middle|\, \bar{\phi} = \Psi^{-1}\Big[\sum_{j=1}^{k} w_j \Psi(\phi^j)\Big]\right)$$

Admixtures/topic models/mixed-membership models:

- LDA [Blei et al. 2003] - An admixture of Multinomials
- Spherical Admixture Model (SAM) [Reisinger et al., 2010] - An admixture of Von-Mises Fisher distributions
- Mixed Membership Stochastic Block Models [Airoldi et al. 2009] - An admixture for generative networks

Background: Poisson MRF

By assuming that the conditional distribution of a variable x_s given all other variables $\mathbf{x}_{\setminus s}$ is a univariate Poisson, a joint Poisson distribution can defined [Yang et al. 2012]:

$$\Pr_{\text{PMRF}}(\mathbf{x} \mid \boldsymbol{\theta}, \Theta) \propto \exp\left\{ \boldsymbol{\theta}^T \mathbf{x} + \mathbf{x}^T \Theta \mathbf{x} - \sum_{s=1}^{p} \ln(x_s!) \right\},$$

where $\boldsymbol{\theta} \in \mathbb{R}^p$ and $\Theta \in \{\mathbb{R}^{p\times p} : \text{diag}(\Theta) = 0\}$.
Node conditionals (i.e. the distribution of one word given all other words) are 1-D Poissons:

$$\Pr(x_s \mid \mathbf{x}_{-s}, \theta_s, \Theta_s) \propto \exp\{ \underbrace{(\theta_s + \mathbf{x}^T \Theta_s)}_{\eta_s} x_s - \ln(x_s!) \}.$$

Background: APM Formal Definition

An Admixture of Poisson MRFs (APM) is an *admixture* with Poisson MRFs as the component distributions:

$$\Pr_{\text{APM}}(\mathbf{x}, \mathbf{w}, \boldsymbol{\theta}^{1\ldots k}, \Theta^{1\ldots k}) =$$

$$\Pr_{\text{PMRF}}\left(\mathbf{x} \,\middle|\, \bar{\theta} = \sum_{j=1}^{k} w_j \theta^j, \bar{\Theta} = \sum_{j=1}^{k} w_j \Theta^j \right) \Pr_{\text{Dir}}(\mathbf{w}) \prod_{j=1}^{k} \Pr(\theta^j, \Theta^j)$$

[Yang et al. 2012] Yang, E., Ravikumar, P., Allen, G. I., and Liu., Z. Graphical Models via Generalized Linear Models. In *NIPS*, 2012.
[Hsieh et al. 2014] Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., and Ravikumar, P. QUIC: Quadratic Approximation for Sparse Inverse Covariance Estimation. *JMLR*, 2014.
[Boyd-Graber et al. 2006] Boyd-Graber, J., Fellbaum, C., Osherson, D., and Schapire, R. Adding Dense, Weighted Connections to {WordNet}. In *Global {WordNet} Conference*, 2006.

Parallel Alternating Newton-like Algorithm (Code available*)

1. Split the algorithm into alternating steps
   - Posterior is convex in W or $(\boldsymbol{\theta}^{1\ldots k}, \Theta^{1\ldots k})$ but not both
   - Similar to EM for mixture models or ALS for NMF

$$\operatorname*{arg\,min}_{\boldsymbol{\Phi}^1, \boldsymbol{\Phi}^2, \cdots, \boldsymbol{\Phi}^p} -\frac{1}{n}\sum_{s=1}^{p}\Big[\text{tr}(\tilde{\mathbf{Z}}^s \boldsymbol{\Phi}^s) - \sum_{i=1}^{n} \exp(\mathbf{z}_i^T \boldsymbol{\Phi}^s \mathbf{w}_i)\Big] + \sum_{s=1}^{p} \lambda \|\text{vec}(\boldsymbol{\Phi}^s)_{\setminus 1}\|_1$$

$$\operatorname*{arg\,min}_{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_n \in \Delta^k} -\frac{1}{n}\sum_{i=1}^{n}\Big[\psi_i^T \mathbf{w}_i - \sum_{s=1}^{p} \exp(\mathbf{z}_i^T \boldsymbol{\Phi}^s \mathbf{w}_i)\Big]$$

where $\mathbf{z}_i = [1\ \mathbf{x}_i^T]^T$, $\phi_s^j = [\theta_s^j\ (\Theta_s^j)^T]^T$, $\boldsymbol{\Phi}^s = [\phi_s^1\, \phi_s^2 \cdots \phi_s^k]$, $\tilde{\mathbf{Z}}^s = f(X, \mathrm{W})$, $\psi_i = f(X, \boldsymbol{\Phi}^{1\ldots k})$

2. Use proximal Newton-like method [Hsieh et al. 2014]
3. Simplify Newton step computation [Hsieh et al. 2014]
   - Compute Hessian entries only for non-zero/free parameters

* Code available at: `http://bigdata.ices.utexas.edu/software/apm/`

Timing Results on Wikipedia

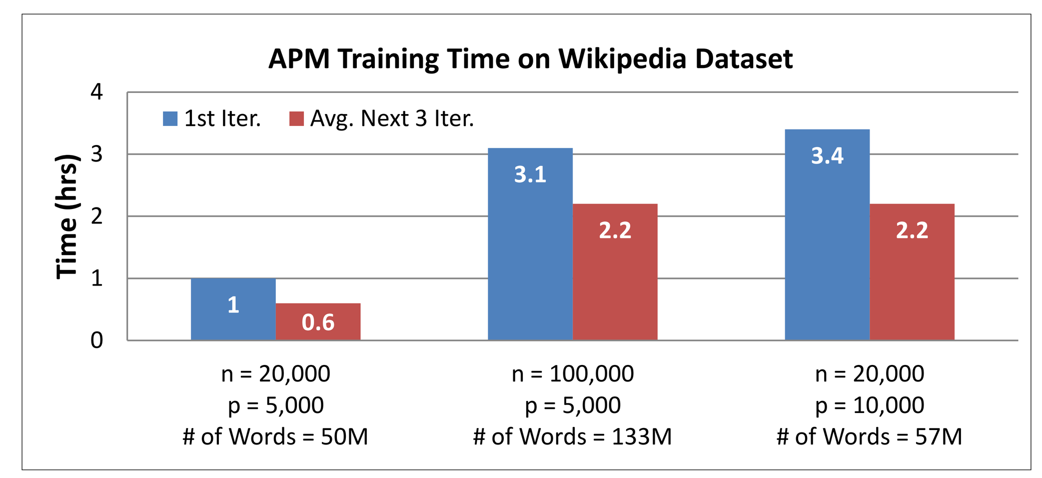


Figure: Timing results for different sizes of a Wikipedia dataset show that the algorithm scales approximately as $O(np^2)$. (Fixing $k = 5, \lambda = 0.5$)

Parallel Speedup on BNC Corpus

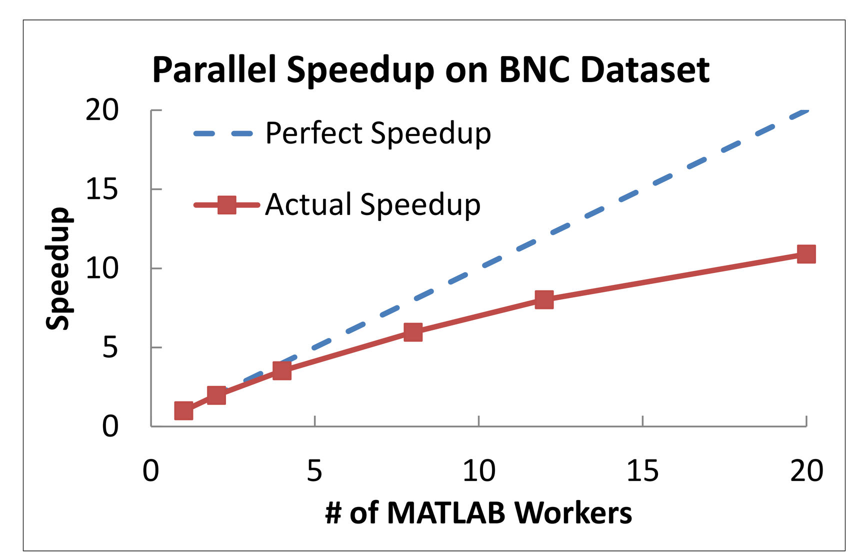


Figure: Parallel speedup is approximately linear when using a simple `parfor` loop in MATLAB. Subproblems are all independent so speedup could be $O(\min(n, p))$ on distributed system. Experiment on BNC corpus ($p = 1646$ and $n = 4049$) fixing $k = 5$, $\lambda = 8$ and running for 30 alternating iterations.

Evocation [Boyd-Graber et al. 2006]

- *Evocation* denotes the idea of which words "evoke" or "bring to mind" other words
- Distinctive from word similarity or synonymy
- Types of *evocation*: Rose - Flower (example), Brave - Noble (kind), Yell - Talk (manner), Eggs - Bacon (co-occurence), Snore - Sleep (setting), Wet - Desert (antonymy), Work - Lazy (exclusivity), Banana - Kiwi (likeness).

Evocation Metric

Rank by model weights $\mathcal{M}$ → Sum top-m human scores $\mathcal{H}$

Word Pair	$\mathcal{H}$	$\mathcal{M}$
w1 ↔ w2	?	0.01
w1 ↔ w3	23	0.1
w1 ↔ w4	0	0.001
w2 ↔ w3	?	12.4
w2 ↔ w4	60	0.67
w3 ↔ w4	5	1.1

Word Pair	$\mathcal{H}$	$\mathcal{M}$
w2 ↔ w3	?	12.4
w3 ↔ w4	5	1.1
w2 ↔ w4	60	0.67
w1 ↔ w3	23	0.1
w1 ↔ w2	?	0.01
w1 ↔ w4	0	0.001

Word Pair	$\mathcal{H}$	$\mathcal{M}$
w3 ↔ w4	5	1.1
w2 ↔ w4	60	0.67
w1 ↔ w3	23	0.1
w1 ↔ w4	0	0.001

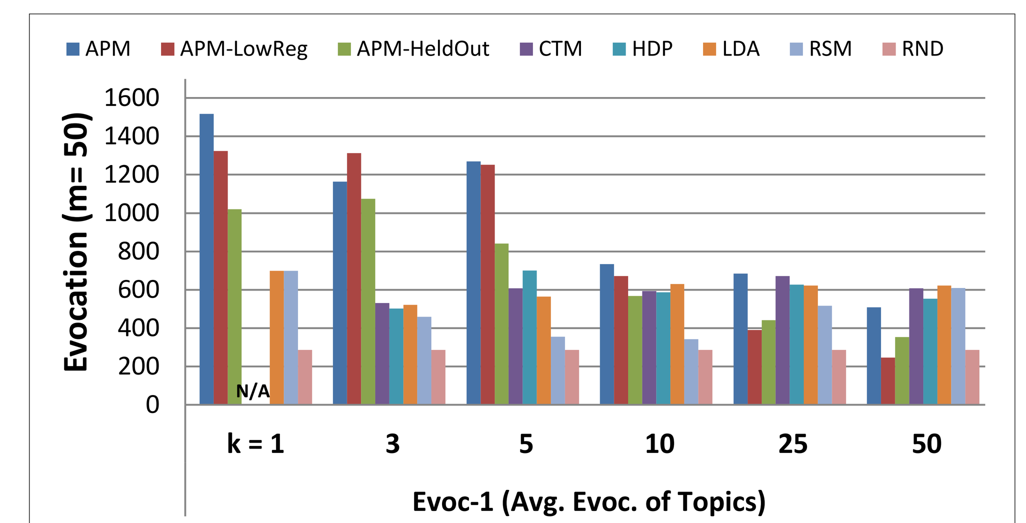
$m =$ Number of top word pairs to evaluate
$\mathcal{H} =$ Human-evaluated scores for subset of word pairs
$\mathcal{M} =$ Corresponding weights induced by model
$\pi_{\mathcal{M}}(j) =$ Ordering induced by $\mathcal{M}$ so that $\mathcal{M}_{\pi(1)} \geq \mathcal{M}_{\pi(2)} \geq \cdots \geq \mathcal{M}_{\pi(|\mathcal{H}|)}$

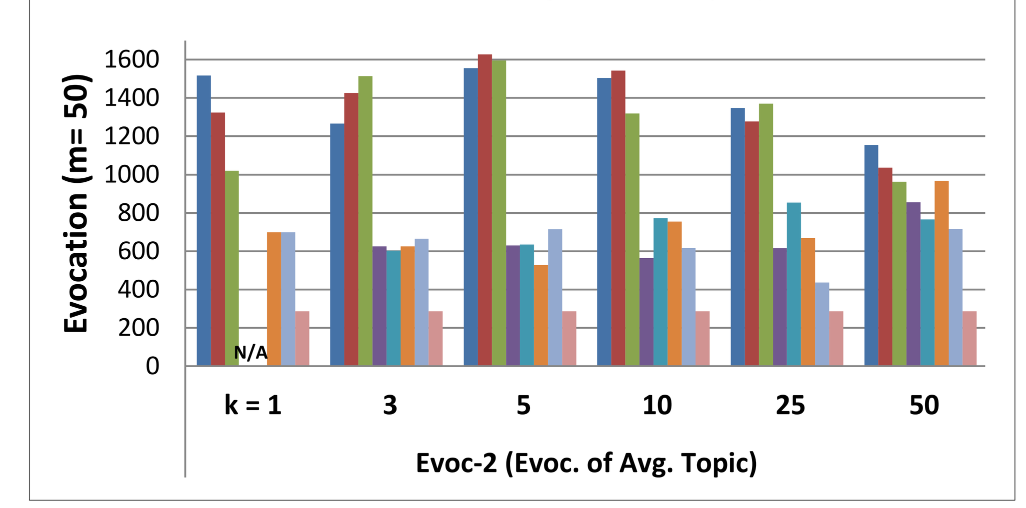
$$\text{Evoc}_m(\mathcal{M}, \mathcal{H}) = \sum_{j=1}^{m} \mathcal{H}_{\pi_{\mathcal{M}}(j)} \qquad \text{(Evocation for Single Topic)}$$

$$\text{Evoc-1} = \sum_{j=1}^{k} \frac{1}{k} \text{Evoc}_m(\mathcal{M}^j, \mathcal{H}) \qquad \text{(Avg. Evoc. of Topics)}$$

$$\text{Evoc-2} = \text{Evoc}_m\Big(\sum_{j=1}^{k} \frac{1}{k} \mathcal{M}^j, \mathcal{H}\Big) \qquad \text{(Evoc. of Avg. Topic)}$$

Evocation Metric Results





Qualitative Analysis of Evocation

- Word pairs for Evoc-2 ($m = 50$) ordered by human score

Best LDA Model ($k = 50$)

Human Score	Word Pair
100	*run.v ↔ car.n*
82	**teach.v ↔ school.n**
69	school.n ↔ class.n
63	*van.n ↔ car.n*
51	hour.n ↔ day.n
50	teach.v ↔ student.n
44	*house.n ↔ government.n*
44	week.n ↔ day.n
38	university.n ↔ institution.n
38	state.n ↔ government.n
38	**woman.n ↔ man.n**
38	give.v ↔ church.n
38	wife.n ↔ man.n
38	**engine.n ↔ car.n**
35	**publish.v ↔ book.n**
32	*west.n ↔ state.n*
32	year.n ↔ day.n
25	*member.n ↔ give.v*
25	dog.n ↔ animal.n
25	*seat.n ↔ car.n*

Best APM Model ($k = 5$)

Human Score	Word Pair
100	**telephone.n ↔ call.n**
97	**husband.n ↔ wife.n**
82	residential.a ↔ home.n
76	politics.n ↔ political.a
75	**steel.n ↔ iron.n**
75	**job.n ↔ employment.n**
75	room.n ↔ bedroom.n
72	**aunt.n ↔ uncle.n**
72	printer.n ↔ print.v
60	love.v ↔ love.n
57	question.n ↔ answer.n
57	prison.n ↔ cell.n
51	**mother.n ↔ baby.n**
50	sun.n ↔ earth.n
50	west.n ↔ east.n
44	**weekend.n ↔ sunday.n**
41	wine.n ↔ drink.v
38	south.n ↔ north.n
38	morning.n ↔ afternoon.n
38	**engine.n ↔ car.n**

- *Red* highlights pairs that seem semantically uninteresting
- Blue highlights pairs that seem semantically interesting

http://www.cs.utexas.edu/~dinouye/
{dinouye,pradeepr,inderjit}@cs.utexas.edu