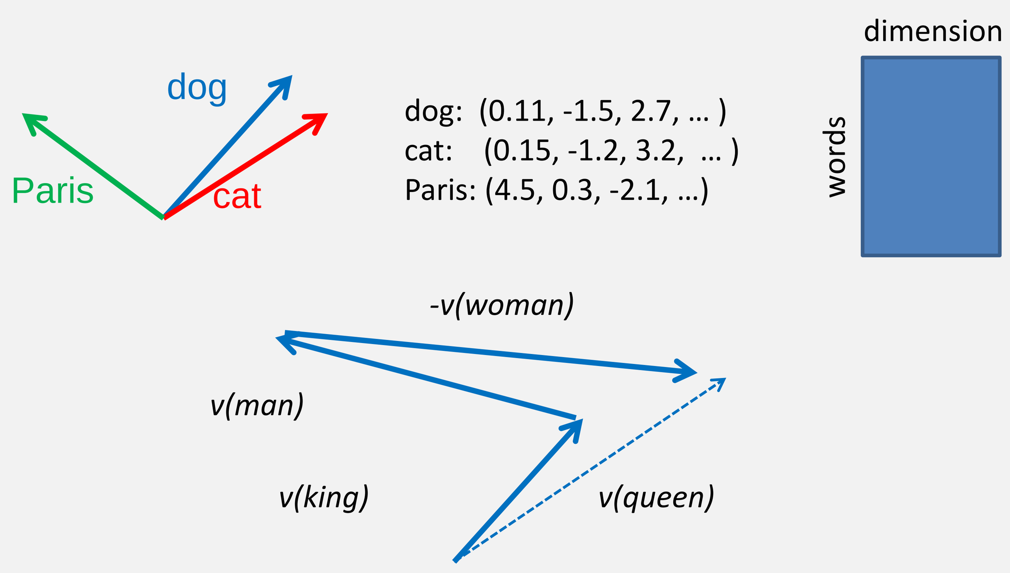
Mixed Membership Word Embeddings: Corpus-Specific Embeddings Without Big Data

James Foulds

University of California, San Diego

Overview

- Word embeddings represent **dictionary words** with **vectors**. Similar words have similar vectors.



Analogy task: man is to king as woman is to _____?

- **Improved performance** for many NLP tasks
  - translation, part-of-speech tagging, chunking, NER, ...
- NLP "from scratch," **without feature engineering**
- Typically trained in **big data** setting

Contributions of this Work

- Demonstrate that **small data** setting is valuable
- Novel embedding model for small data setting, leveraging connections to **topic models**
  - **Mixed membership** representation for parameter sharing
- **Efficient training**, using recent advances from both topic models and word embeddings
  - **Metropolis-Hastings-Walker** algorithm (Li et al., 2014)
  - **Noise-contrastive estimation** (Gutmann and Hyvarinen, 2010, 2012)
- Experimental study; practical recommendations

Background

Skip-gram model (Mikolov et al., 2013)

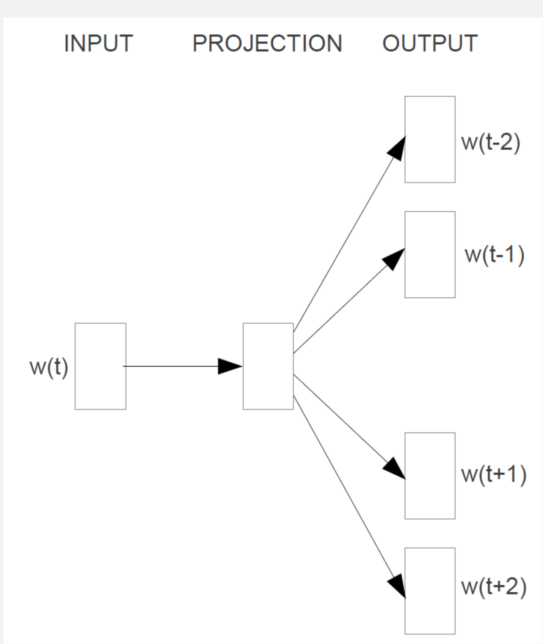


Figure due to Mikolov et al. (2013)

A log-bilinear classifier for the **context of a given word**

$$p(w_j|w_i) \propto \exp(v'^{\intercal}_{w_j} v_{w_i} + b_j)$$

v_w : "input" vectors

v'_w : "output" vectors

b_j : bias term

- Simple model scales to large data sets
  - Beats deep neural network models

Noise-contrastive estimation (NCE)

(Gutmann and Hyvarinen, 2010, 2012; Mnih & Teh, 2012)

- Train a logistic regression classifier to distinguish between data and noise samples

$$J^{w_i}(\theta) = E_{p_{data}}[\log p(D=1|w_j, w_i, \theta)] + kE_{p_{noise}}[\log p(D=0|w_j, w_i, \theta)]$$

$$p(D=1|w_j, w_i, \theta) = \frac{p_\theta^{w_i}(w_j)}{p_\theta^{w_i}(w_j) + kp_{noise}(w_j)}$$ (*D* = *1* if data, *D=0* if noise)

- Sublinear in vocab size *V*, unlike MLE
- Linear in # samples, independent of *V*
- Approaches MLE as # samples *k* increases

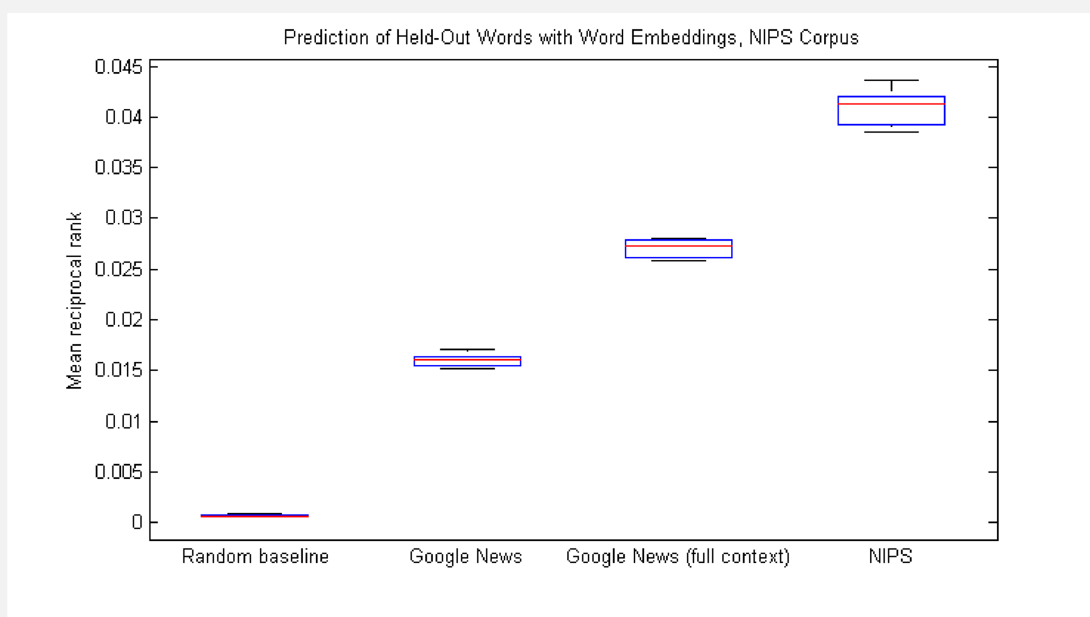
Connections to Topic Models. Mixed Membership Extension to the Skip-Gram

- Skip-gram corresponds to a **supervised naïve Bayes topic model**, up to its parameterization via embeddings
- I propose topic model and mixed membership variants
- **Mixed membership** models provide parameter sharing
- Can use **fewer vectors than words** for small data, while retaining substantial representational power

	Skip-gram	Skip-gram topic model
Naive Bayes	For each word in the corpus w_i For each word $w_j \in context(i)$ Draw $w_j \vert w_i$ via $p(w_j \vert w_i) \propto exp(v'^{\intercal}_{w_j} v_{w_i} + b_j)$	For each word in the corpus w_i For each word $w_j \in context(i)$ Draw $w_j \vert w_i \sim \text{Discrete}(\phi^{(w_i)})$
Mixed membership	For each word in the corpus w_i Draw a topic $z_i \sim \text{Discrete}(\theta^{(w_i)})$ For each word $w_j \in context(i)$ Draw $w_j \vert w_i$ via $p(w_j \vert w_i) \propto exp(v'^{\intercal}_{w_j} v_{z_i} + b_j)$	For each word in the corpus w_i Draw a topic $z_i \sim \text{Discrete}(\theta^{(w_i)})$ For each word $w_j \in context(i)$ Draw $w_j \vert w_i \sim \text{Discrete}(\phi^{(z_i)})$

The Case for Small Data

- Many (most?) data sets of interest are **small**
  - E.g. NIPS corpus, 1740 articles
- Common practice: use vectors trained on **another, larger corpus**
  - Tomas Mikolov's vectors from Google News, 100B words
  - Wall Street Journal corpus



- Similar words to "*learning*" based on different corpora:
  - **Google News:** teaching learn Learning reteaching learner_centered emergent_literacy kinesthetic_learning teach
  - **NIPS:** reinforcement belief learning policy algorithms Singh robot machine MDP planning algorithm problem methods function
- Word embeddings **biased** by their training dataset, **no matter how large**. E.g. can encode **sexist assumptions** (Bolukbasi et al., 2016)

"***man*** is to ***computer programmer*** as ***woman*** is to ***homemaker***"

Inference for MM Skip-Gram Topic Model

- Bayesian inference w/ Dirichlet priors, collapsed Gibbs sampling

$$p(z_i = k|\cdot) \propto \left(n_k^{(w_i)\neg i} + \alpha_k\right) \prod_{c=1}^{|\text{context}(i)|} \frac{n_{w_c^{(i)}}^{(k)\neg i} + \beta_{w_c^{(i)}} + n_{w_c^{(i)}}^{(i,c)}}{n^{(k)\neg i} + \sum_{w'} \beta_{w'} + c - 1}$$

- Scale to many topics: Metropolis-Hastings-Walker
- Alias table data structure, amortized O(1) sampling
- "Mixture of experts" proposal, alias tables for words

$$q(k) = \sum_{c=1}^{|\text{context}(w_i)|} \frac{1}{|\text{context}|} q_{w_c^{(i)}}(k) \ , \ q_{w_c^{(i)}}(k) = \frac{1}{Z_{w_c}} \alpha_k \frac{n_{w_c^{(i)}}^{(k)} + \beta_{w_c^{(i)}}}{n^{(k)} + \sum_{w'} \beta_{w'}}$$

- Simulated annealing to escape early local maxima

Approximate MLE for MM Skip-Gram

- Online EM impractical
  - M-step is O(V), E-step is O(KV)
- Approximate online EM
  - Key insight: MMSG topic model **equivalent** to word embedding model, *up to the Dirichlet prior*
    - **Pre-solve E-step** via topic model CGS algorithm
    - Apply **NCE** to solve M-step
  - Overall algorithm approximates maximum likelihood estimation via these two principled approximations

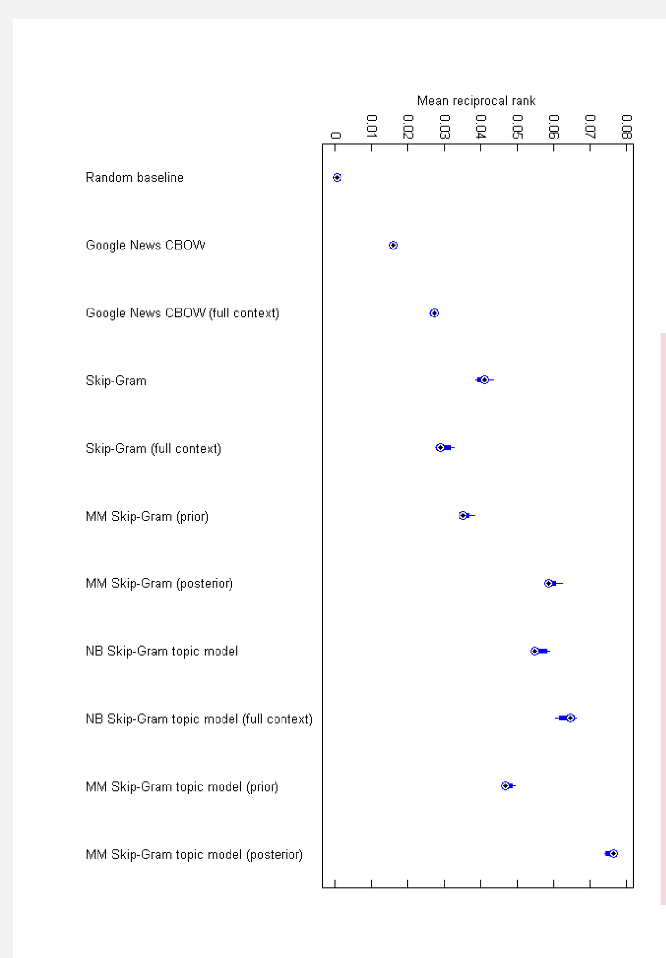
Experimental Results: NIPS Corpus

Input word = "Bayesian"

Model	Top words in topic for input word. Top 3 topics for word shown for mixed membership models.
SGTM	model networks learning neural bayesian data models approach network framework
SG	belief learning framework models methods markov function bayesian based inference
MMSGTM	bayesian model parameters posterior prior distribution approach likelihood variational inference neural networks computation bayesian learning mackay framework network functions practical carlo monte bayesian gaussian neural neal implementation methods models williams
MMSG	variational likelihood bayesian inference approach parameters marginal dirichlet posterior sampling neural bayesian learning networks computation framework regularization entropy press mackay neal rasmussen monte bayesian models http press neural barber carlo

Input word = "Jordan"

Model	Top words in topic for input word. Top 3 topics for word shown for mixed membership models.
SGTM	neural learning jacobs jordan algorithm experts mit models em networks
SG	jacobs rumelhart mozer petsche jaakkola nowlan jordan supervised learning michael
MMSGTM	experts mixtures jordan neural jacobs hinton computation local em nowlan jordan models learning graphical mit jaakkola press psyche saul ghahramani neural information processing advances systems mit press editors cambridge touretzky
MMSG	mixtures experts jacobs hierarchical nowlan neal hinton press em adaptive pages press mit graphical kluwer variational jaakkola learning saul models press mit pages information processing neural advances reinforcement eds learning



Prediction task:

- Predict context words, given an input word.
- Treat as ranking problem, mean reciprocal rank metric

Using the **full context** (posterior over topic or summing vectors) **helps all models except the basic skip-gram**

Mixed-membership models (w/ posterior) beat naïve Bayes models, for both word embedding and topic models

Topic models beat their corresponding embedding models, for both **naïve Bayes** and **Mixed Membership** models

jfoulds@ucsd.edu