UC San Diego

QUALCOMM INSTITUTE

Mixed membership word embeddings:

Corpus-specific embeddings **without big data**

James Foulds

University of California, San Diego

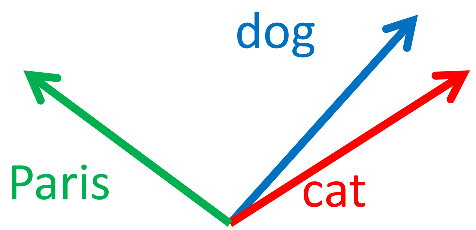
Southern California Machine Learning Symposium, Caltech, 11/18/2018

UCSDCSE
Computer Science and Engineering

Calit2

Word Embeddings

- Language models which learn to represent **dictionary words** with **vectors**

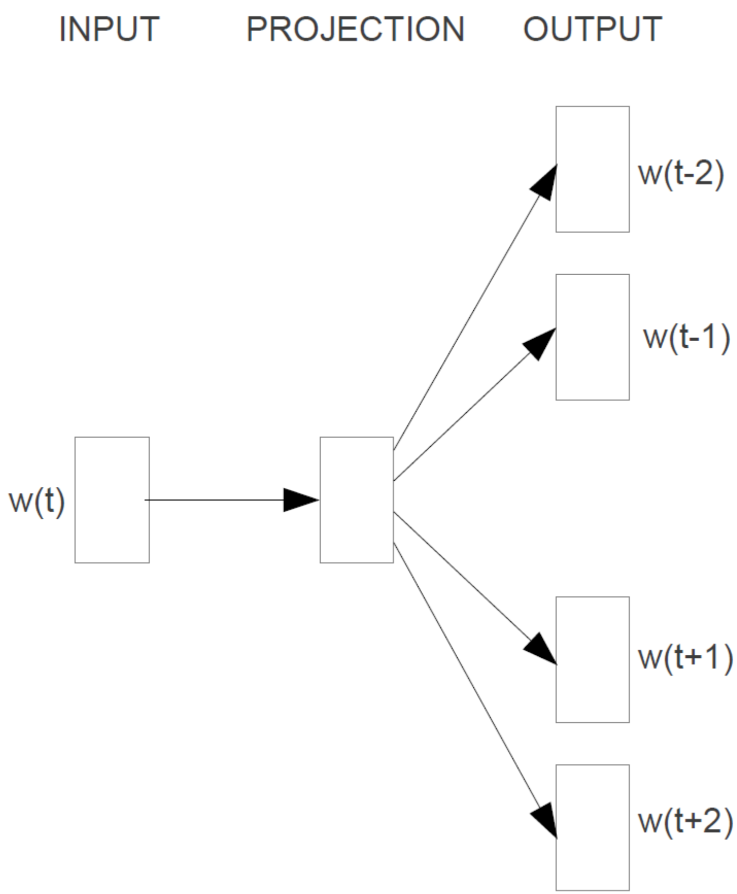


dog: (0.11, -1.5, 2.7, …)
cat: (0.15, -1.2, 3.2, …)
Paris: (4.5, 0.3, -2.1, …)

- Nuanced representations for words
- Improved performance for many NLP tasks
  - translation, part-of-speech tagging, chunking, NER, …
- NLP "from scratch"? (Collobert et al., 2011)

Word2vec (Mikolov et al., 2013)

Skip-Gram



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

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

$$v_w : \text{“input” vectors}$$

$$v'_w : \text{“output” vectors}$$

Figure due to Mikolov et al. (2013)

Word2vec (Mikolov et al., 2013)

- Key insights:
  - **Simple models** can be trained efficiently on **big data**
  - High-dimensional simple embedding models,
    trained on massive data sets,
    can **outperform sophisticated neural nets**

Target Corpus vs Big Data?

- Suppose you want word embeddings to use on the NIPS corpus, 1740 docs

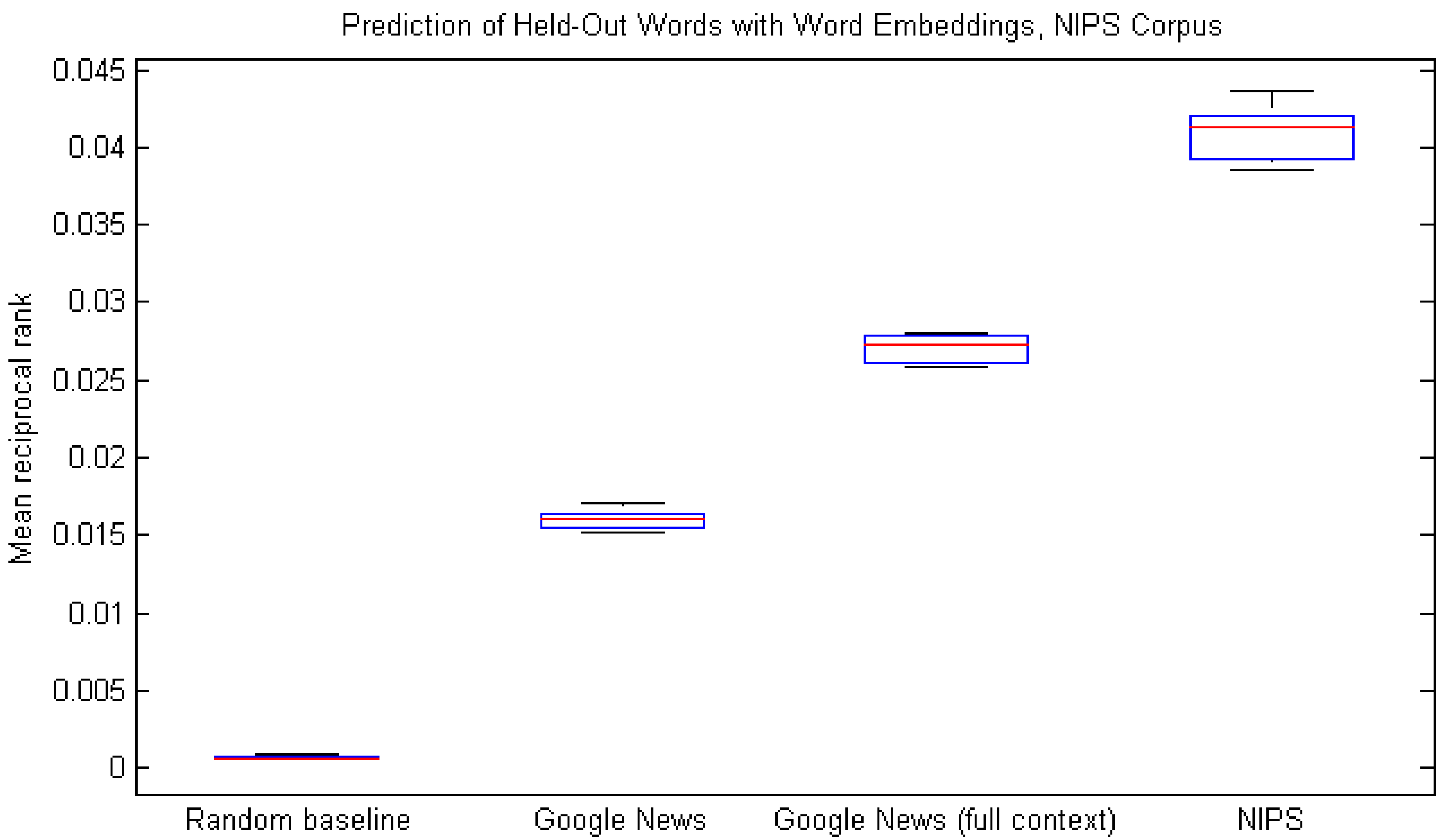
  Which has better predictive performance for held out word/context-word pairs on NIPS corpus?

  - **Option 1:** Word embeddings **trained on NIPS**.
    2.3 million word tokens, 128 dim vectors

  - **Option 2:** embeddings trained on **Google News**.
    100 billion word tokens, 300 dim vectors

Target Corpus vs Big Data?

- **Answer: Option 1**, embeddings trained on NIPS



Similar Words to "*learning*" for each Corpus

- **Google News:** teaching learn Learning reteaching learner_centered emergent_literacy kinesthetic_learning teach learners learing lifeskills learner experiential_learning Teaching unlearning numeracy_literacy taught cross_curricular Kumon_Method ESL_FSL

- **NIPS:** reinforcement belief learning policy algorithms Singh robot machine MDP planning algorithm problem methods function approximation POMDP gradient markov approach based

The Case for Small Data

- Many (most?) data sets of interest are **small**
  - E.g. NIPS corpus, 1740 articles

- Common practice:
  - Use word vectors trained on **another, larger corpus**
    - Tomas Mikolov's vectors from Google News, 100B words
    - Wall Street Journal corpus

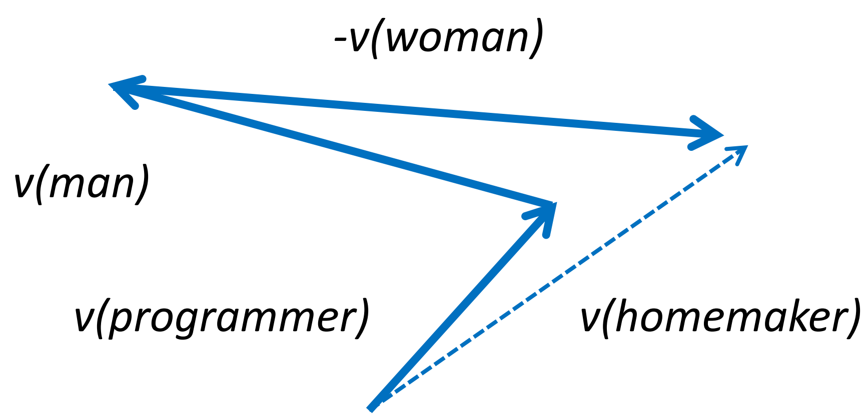
- In many cases, this may not be the best idea

The Case for Small Data

- Word embedding models are **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***"



The Case for Small Data

- Although powerful,
  big data will not solve all our problems!

- We still need effective quantitative methods
  for small data sets!

Contributions

- **Novel model for word embeddings on small data**
  - parameter sharing via mixed membership
- **Efficient training algorithm**
  - Leveraging advances in word embeddings (NCE) and topic models (Metropolis-Hastings-Walker)
- **Empirical study**
  - Practical recommendations

The Skip-Gram as a Probabilistic Model

- Can view skip-gram as probabilistic model for ``generating'' context words

For each word in the corpus w_i

For each word $w_j \in context(i)$

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

Implements **distributional hypothesis**

Conditional discrete distribution over words: can identify with a **topic**

The Skip-Gram as a Probabilistic Model

Naïve Bayes conditional independence

For each word in the corpus w_i

For each word $w_j \in context(i)$

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

Observed "**cluster**" assignment

"**Topic**" distribution for input word w_i

Mixed Membership Modeling

- Naïve Bayes conditional independence assumption typically too strong, not realistic

- Mixed membership: relax “hard clustering” assumption to “soft clustering”
  - Membership distribution over clusters
    E.g.:
    - Text documents belong to a distribution of topics
    - Social network individuals belong partly to multiple communities

Grid of Models' "Generative" Processes

Identifying word distributions with topics leads to analogous topic model

	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 $w_j \| w_i$ via $p(w_j \| w_i) \propto exp({v'_{w_j}}^\intercal v_{w_i} + b_j)$	For each word in the corpus w_i For each word $w_j \in context(i)$ Draw $w_j \| w_i \sim \text{Discrete}(\phi^{(w_i)})$
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 $w_j \| w_i$ via $p(w_j \| w_i) \propto exp({v'_{w_j}}^\intercal v_{z_i} + b_j)$	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 $w_j \| w_i \sim \text{Discrete}(\phi^{(z_i)})$

Reinstate word vector representation

Relax naïve Bayes assumption, replace with mixed membership model.
- -flexible representation for words
- -parameter sharing

Mixed Membership Skip-Gram Posterior Inference for Topic Vector

- **Context** can be leveraged for inferring the topic vector at test time, via **Bayes' rule**:

$$Pr(v_{w_i} = v_k | w_i, \text{context}(i), \mathbf{V}, \Theta) \propto Pr(z_i = k | w_i, \Theta) Pr(\text{context}(i) | z_i = k, \mathbf{V})$$

$$= \theta_k^{(w_i)} \prod_{c \in \text{context}(i)} \frac{exp(v'^{\intercal}_{w_c^{(i)}} v_k)}{\sum_{j'=1}^{V} exp(v'^{\intercal}_{j'} v_k)}$$

Bayesian Inference for MMSG Topic Model

- Bayesian version of model with Dirichlet priors
- Collapsed Gibbs sampling

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

Bayesian Inference for MMSG Topic Model

- Challenge 1: want relatively large # topics
- Solution: **Metropolis-Hastings-Walker** algorithm (Li et al. 2014)
  - **Alias table** data structure, amortized O(1) sampling
  - **Sparse implementation**, sublinear in topics K
  - **Metropolis-Hastings correction** for sampling from stale distributions

Metropolis-Hastings-Walker
(Li et al. 2014)

Sparse

Dense, slow-changing

$$p(z_i = k|\cdot) \propto n_k^{(w_i)\neg i} A_{ik} + \alpha_k A_{ik}$$

$$A_{ik} = \prod_{c=1}^{|\text{context}(i)|} \frac{n_{w_c^{(i)}}^{(k)\neg i} + \beta_{w_c^{(i)}} + n_{w_c^{(i)}}^{(i,c)}}{n^{(k)\neg i} + \sum_{w'} \beta_{w'} + c - 1}$$

- Approximate second term of the mixture, sample efficiently via alias tables, correct via Metropolis

Metropolis-Hastings-Walker Proposal

- Dense part of Gibbs update is a "*product of experts*" (Hinton, 2004), expert for each context word

- Use a "*mixture of experts*" proposal distribution

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

- Can sample efficiently from "experts" via alias tables

Bayesian Inference for MMSG Topic Model

- Challenge 2: cluster assignment updates almost deterministic, vulnerable to local maxima

- Solution: **simulated annealing**
  - Anneal temperature of model
    - adjusting Metropolis-Hastings acceptance probabilities

Approximate MLE for Mixed Membership 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 Dirichlet prior
    - **Pre-solve E-step** via topic model CGS
    - Apply **Noise Contrastive Estimation** to solve M-step
  - Entire algorithm approximates maximum likelihood estimation via these two principled approximations

Qualitative 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

Qualitative 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

Qualitative 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

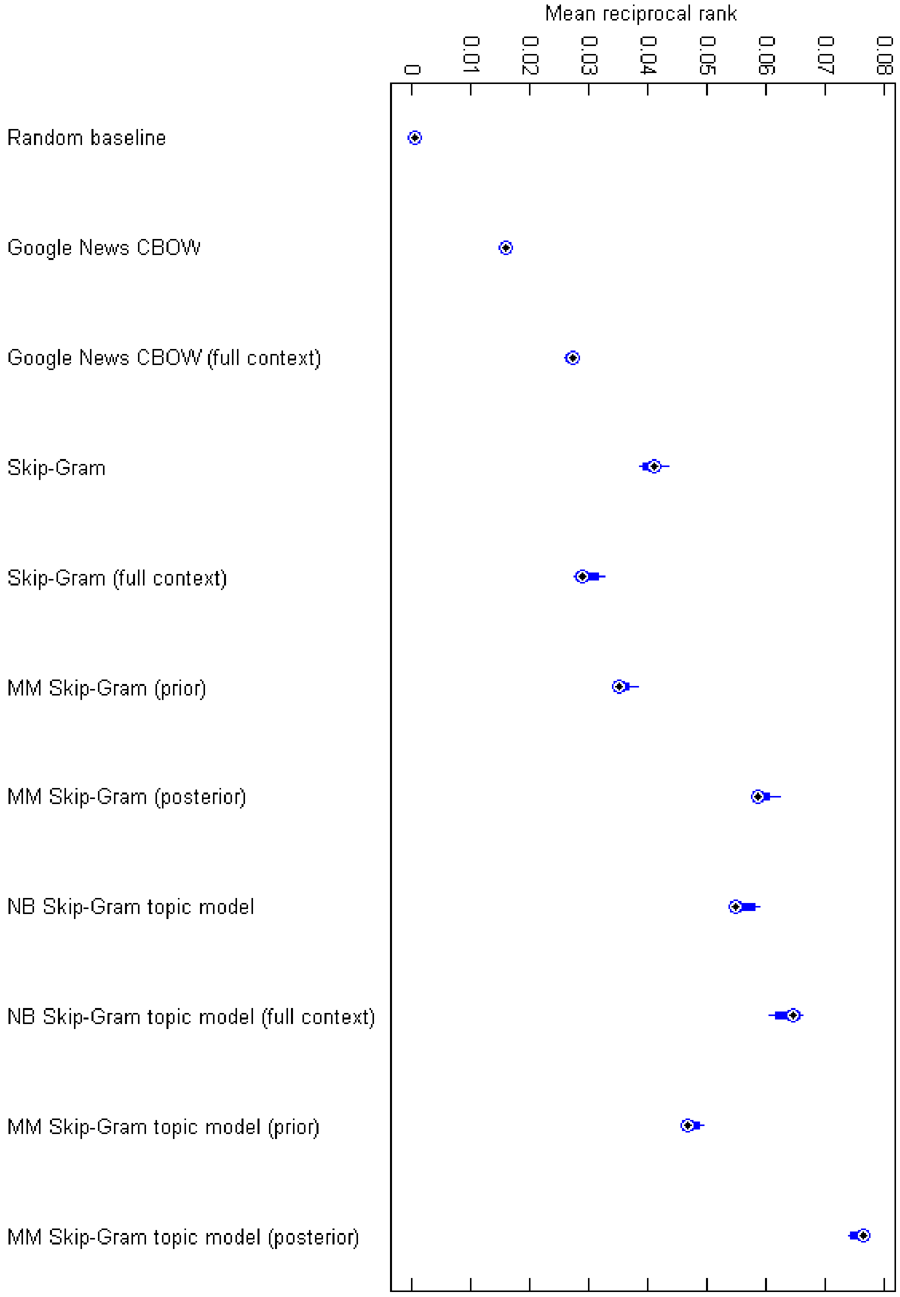
Qualitative Results, NIPS Corpus

	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

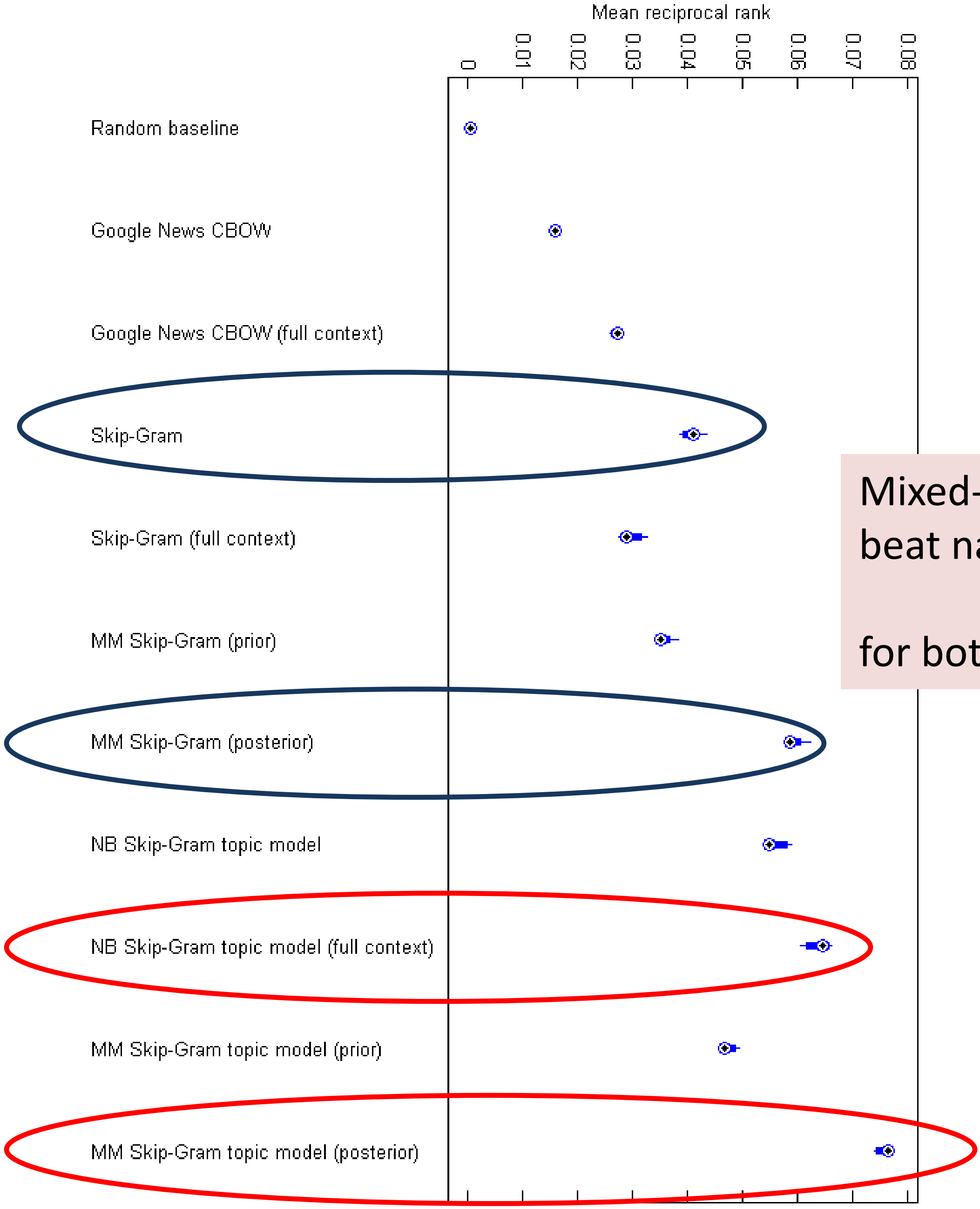
Qualitative Results, NIPS Corpus

	Input word = "SVM"
Model	Top words in topic for input word. Top 3 topics for word shown for mixed membership models.
SGTM	svm algorithm training method set support vector kernel data error
SG	svm svms performance smo results figure learning algorithms function problem
MMSGTM	method svm parzen figure probability shows distribution gaussians mixture density
	smo kernel svm chunking wij light time linear sparse faster
	data kernel vector support class set vectors training estimate function
MMSG	parzen svm pact xll method xla forty ibr substitution figure
	smo svm advantage numerical speed light terms support estimator kernel
	function support vector svm vectors relevance class svms working kernel

Prediction Performance, NIPS Corpus



Prediction Performance, NIPS Corpus



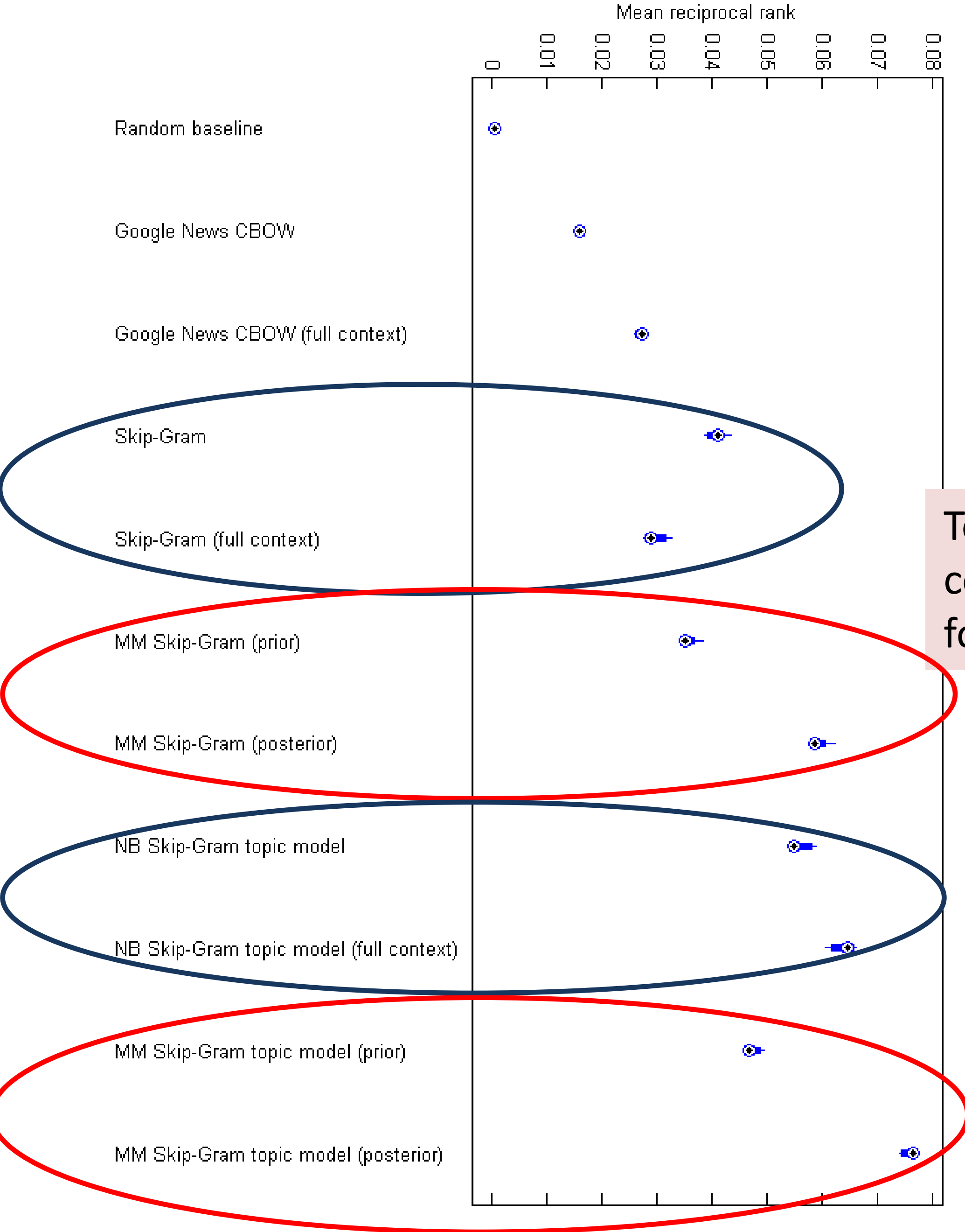
Mixed-membership models (w/ posterior) beat naïve Bayes models,

for both word embedding and topic models

Prediction Performance, NIPS Corpus

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

Prediction Performance, NIPS Corpus



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

Open question: when do we really need word vector representations???

Conclusion

- **Small data** still matters!!
- Proposed **mixed membership**, **topic model** versions of skip-gram word embedding models
- **Efficient training** via MHW collapsed Gibbs + NCE
- Proposed models **improve prediction**
- Ongoing/future work:
  - Evaluation on more **datasets**, **downstream tasks**
  - Adapt to **big data** setting as well?