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

James Foulds
University of California, San Diego

Abstract

Word embeddings provide a nuanced representation of words which can improve the performance of NLP systems by revealing the hidden structural properties of words and their relationships to each other. These models have recently risen in popularity due to the successful performance of scalable algorithms trained in the big data setting. Consequently, word embeddings are commonly trained on very large generic corpora such as Wikipedia, instead of a (typically smaller) corpus of interest, leading to embeddings that are precisely trained but inaccurate relative to the domain of study. I propose a probabilistic model-based word embedding method which can recover high-dimensional embeddings without big data, due to a sensible parameter sharing scheme. The key insight is to leverage the notion of *mixed membership* modeling, in which global representations are shared, but individual entities (i.e. dictionary words) are free to use these representations to uniquely differing degrees. Leveraging connections to topic models, I show how to train these models in high dimensions using a combination of state-of-the-art training techniques from the word embedding and topic modeling literatures.

1 Introduction and Background

Traditional language models aim to predict words given the contexts that they are found in, thereby forming a joint probabilistic model for sequences of words in a language. Bengio et al. (2003) developed improved language models by using *distributed representations* (Hinton et al., 1986), in which words are represented by neural network synapse weights, or equivalently, vector space embeddings. More recently, these latent *word embeddings* have been shown to be valuable for other downstream NLP tasks such as statistical machine translation (Vaswani et al., 2013), part-of-speech tagging, chunking, and named entity recogition (Collobert et al., 2011), as they provide a more nuanced representation of words than a simple indicator vector into a dictionary.

Word embeddings have risen in popularity for NLP applications due to the success of simplified language models designed specifically for learning embeddings and which are trained in the big data setting. In particular, Mikolov et al. (2013a,b) proposed the *skip-gram* model, which inverts the language model prediction task and aims to *predict the context* given an input word. The skip-gram model is a log-bilinear probabilistic classifier parameterized by "input" word embedding vectors v_{w_i} for the input words w_i, and "output" word embedding vectors v'_{w_j} for context words $w_j \in \text{context}(i)$. Although the skip-gram is discriminative as it does not jointly model the input words w_i, in this work I equivalently interpret it as encoding a "generative" process for the context given the words (Table 1, top-left), in order to develop probabilistic models that extend the skip-gram.

2 The Mixed Membership Skip-Gram

The main insight from the work of (Mikolov et al., 2013a,b) is that simple word embedding models with high-dimensional representations can scale up to large datasets, allowing them to outperform

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

Table 1: "Generative" processes. Identifying the skip-gram (top-left)'s word distributions with topics yields analogous topic models (right), and mixed membership modeling extensions (bottom).

more sophisticated neural network models. The standard practice is to train these models on large corpora such as Wikipedia, and use the embeddings for NLP tasks on other smaller datasets. However, even large corpora have idiosyncrasies that may make their embeddings invalid for other domains. Word embeddings from standard corpora can encode sexist assumptions (Bolukbasi et al., 2016), and it is reasonable to expect that they also encode the dominant white male Eurocentric wordview, inappropriate for studying, e.g., black female hip-hop artists' lyrics, or poetry by Syrian refugees. In this work, I propose a model that can be trained directly on a corpus of interest, without the need for a separate big data training set, due to parameter sharing via mixed membership modeling.

To accomplish this, I adapt advances from topic modeling (Blei et al., 2003) (background information omitted for space). Following the *distributional hypothesis* (Harris, 1954), the skip-gram's word embeddings parameterize discrete probability distributions over words $p(w_j|w_i)$ which tend to co-occur, and tend to be semantically coherent – a property leveraged by Gaussian LDA (Das et al., 2015). By identifying these discrete distributions with *topics* $\phi^{(w_i)}$, we see that the skip-gram corresponds to a supervised naive Bayes topic model, where input words w_i are observed "cluster assignments," the words in all of w_i's contexts are a "document," and "topics" are parameterized by word vectors (Table 1, top-right). LDA topic models improve over naive Bayes by using a *mixed membership* model, in which the assumption that all words in a document belong to the same topic is relaxed, and replaced with a *distribution* over topics. By applying the mixed membership representation to the topic model version of the skip-gram, we obtain the model in the bottom-right of Table 1.[1] After once again parameterizing this model with word embeddings, we obtain our final model, the *mixed membership skip-gram* (Table 1, bottom-left). In the model, each input word has a distribution over topics $\theta^{(w_i)}$. Each topic has a vector-space embedding v_k and each output word has an embedding v'_{w_j}. A topic is drawn for each context, and the words in the context are drawn from the log-bilinear model.

To train the mixed membership skip-gram, an EM algorithm with stochastic gradient M-step updates can readily be derived, with updates similar to those of Tian et al. (2014)'s multi-prototype embedding model. However, this algorithm is impractical due to a $O(KV)$ complexity for the E-step, where K and V are the number of topics/dictionary words, respectively. Instead, I propose to leverage the relationship to the topic model (Table 1, bottom-right), which admits a collapsed Gibbs sampler when Dirichlet priors are used, to solve the E-step via the topic model as a pre-processing step. I have scaled this model up to tens of thousands of topics using an adapted version of the recently proposed Metropolis-Hastings-Walker algorithm for high-dimensional topic models (Li et al., 2014), details omitted for space. Finally, with the E-step and θ computed via the topic model, the noise-contrastive estimation (NCE) algorithm (Gutmann and Hyvärinen, 2012; Mnih and Kavukcuoglu, 2013) can be used to recover the word vectors with a sub-linear dependence on K and V (omitted for space).

3 Conclusion

I have proposed a model-based method for training corpus-specific word embeddings without big data, leveraging mixed membership representations, the Metropolis-Hastings-Walker algorithm and noise-contrastive estimation. Preliminary results (Appendix A) indicate that high-quality embeddings and topics can be obtained using this algorithm. I am currently in the process of rigorously evaluating the method with comparison to strong baselines, and will present updated results in the symposium.

[1] The model retains a naive Bayes assumption at the context level, for latent variable count parsimony.

Acknowledgments: I thank Eric Nalisnick and Padhraic Smyth for many helpful discussions.

References

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3(Feb):1137–1155.

Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *ArXiv preprint arXiv:1607.06520*.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Das, R., Zaheer, M., and Dyer, C. (2015). Gaussian LDA for topic models with word embeddings. In *ACL*, pages 795–804.

Gutmann, M. U. and Hyvärinen, A. (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13(Feb):307–361.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Hinton, G. E., Mcclelland, J. L., and Rumelhart, D. E. (1986). Distributed representations. In *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations*, chapter 3, pages 77–109.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Li, A. Q., Ahmed, A., Ravi, S., and Smola, A. J. (2014). Reducing the sampling complexity of topic models. In *KDD*, pages 891–900. ACM.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *ICLR*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Mnih, A. and Kavukcuoglu, K. (2013). Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in Neural Information Processing Systems*, pages 2265–2273.

Tian, F., Dai, H., Bian, J., Gao, B., Zhang, R., Chen, E., and Liu, T.-Y. (2014). A probabilistic model for learning multi-prototype word embeddings. In *COLING*, pages 151–160.

Vaswani, A., Zhao, Y., Fossum, V., and Chiang, D. (2013). Decoding with large-scale neural language models improves translation. In *EMNLP*, pages 1387–1392.

A Preliminary Results

Example topics from the NIPS corpus are provided in Tables 2 – 4. I used a context window of 5 words before and after the input word, 128-dimensional embeddings, and 2000 topics for the mixed membership models. Note that the (naive Bayes) skip-gram and mixed membership skip-gram aim to learn embeddings that encode the same "topics" as their corresponding topic model variants (up to the Dirichlet prior). In the tables, similarities between the topic models and embedding models' topics are indicative that the NCE algorithm has recovered embeddings that encode the empirical conditional word distributions. On the other hand, differences between embedding model topics and topic model topics are due to either convergence, or to the limitations of the embedding's representational power.

The results show that both the topic models and embedding models can learn interpretable topics. While there were differences between the topic models and word embeddings' word distributions, there was also substantial overlap. The topics for the naive Bayes variants (including the original skip-gram) typically contained words from several different specific types of contexts, while the mixed membership models were able to separate these types of context and represent each of them with their own topics. For the input word "Bayesian" (Table 2), the naive Bayes and skip-gram

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

Table 2: SG = skip-gram, TM = topic model, MM = mixed membership.

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

Table 3: The 5th most likely topic for "Jordan" shown instead of the 3rd, for interest.

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

Table 4: PaCT refers to "*Plug-in Classification Technique*."

models learned a topic with words that refer to Bayesian networks, probabilistic models, and neural networks. The mixed membership models are able to separate this into more coherent and specific topics including Bayesian inference, Bayesian training of neural networks (for which Sir David MacKay was a strong proponent), and Monte Carlo methods (championed by Radford Neal, as well as Carl Rasmussen, Chris Williams, and David Barber, in the context of Gaussian Processes).

The input word "Jordan" (Table 3) refers to Michael I. Jordan, who has played an important role in the NIPS community, both as a researcher and in service activities, including Program Chair, General Chair, editor, and board member. The names in the topics refer to his co-authors and co-editors. In this context, "mit" refers to either his home institution from 1988–1998, or to *MIT press*, the publisher of the NIPS proceedings. The mixed membership models identified distinct topics for his well-known work on mixtures of experts models (including the names of his co-authors of (Jacobs et al., 1991)), graphical models, and a *NIPS conference* topic relating to his service roles within the conference.

For the input word "SVM" (Table 4), the skip-gram and its topic model variant both learned coherent topics relating to support vector machines. The mixed membership models were however able to recover more nuanced topics, including a general data mining topic, a topic on SVM usage which references the popular *SVM light* implementation, and a topic on the algorithmic details of SVMs.