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.
- 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 Hyvärinen, 2010, 2012)
- Experimental study: practical recommendations

Background

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

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

\[ p(w_j | w_i) \propto \exp(v^T_{w_j} v_{w_i} + b_j) \]

- Similar words to “learning” based on different corpora:
  - Google News: teaching learn Language Retraining learner-centered emergent literacy kinesesthetic learning teach
  - NIPS: reinforcement learning policy algorithms Singh robot machine MDP planning algorithm problem methods function

Noise-contrastive estimation (NCE)
(Gutmann and Hyvärinen, 2010, 2012)

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

\[ p(\theta) = \prod_{D \in \text{data}} \log p(D) - \sum_{i=1}^{K} \sum_{w \in V} \log p_{\theta}(w) \]

- Sublinear in vocab size \( V \), unlike MLE
- Linear in \( K \) samples, independent of \( V \)
- Approaches MLE as \( K \) samples \( k \) increases

Inference for MM Skip-Gram Topic Model

- Bayesian inference w/ Dirichlet priors, collapsed Gibbs sampling

\[ p(z_i = k) \propto \left( n_{k,w}^{(i)} + \alpha_k \right) \prod_{w \in V} \sum_{k'} n_{k',w}^{(i)} + \sum_k \beta_k + c - 1 \]

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

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

- Skip-gram corresponds to a supervised naive 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

Naive Bayes
For each word \( w_j \) in context \( i \)
Draw \( w_j | w_i \) via
\[ p(w_j | w_i) \propto \exp(v^T_{w_j} v_{w_i} + b_j) \]

Mixed membership
For each word \( w_j \) in context \( i \)
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
\[ p(w_j | w_i) \propto \exp(v^T_{w_j} \sum_k \beta_k \theta^{(w_i)} + b_j) \]

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

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

- Top words in topic for input word. Top 3 topics for word drawn for mixed membership models.
- SGGTM: mixed membership topic model, small data, no big data
- MGBGM: mixed membership topic model, small data, big data
- Approximate EM: approximate EM with posterior Dirichlet smoothing

The table above demonstrates the performance of the approximate EM algorithm compared to the exact EM algorithm. The approximate EM algorithm achieves similar performance to the exact EM algorithm while reducing computational time.

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 naive Bayes models, for both word embedding and topic models

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

References


Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable feedback and suggestions. This research was supported by a grant from the National Science Foundation (Award Number: 1527294).