Supplemental Materials: Topic modeling with Latent Dirichlet Allocation

Background

There is a rather different class of DS models that has been used to model free association: topic models, and especially latent Dirichlet allocation (LDA; Blei et al., 2003; Griffiths et al., 2007). Like the other DS models that we discussed in the main text, LDA also involves exploiting lexical distributional statistics in large corpora. However, in LDA a document is assumed to have a distribution over various topics, each of which is a distribution over words. Fitting an LDA model to a corpus is thus a matter of finding – with techniques from Bayesian inference – each document's distribution over topics, and each topic's distribution over words. It has been shown that LDA can be especially useful for modeling various aspects of semantic cognition, including free association (Griffiths et al., 2007; Nematzadeh et al., 2017). Importantly, measures derived from LDA can account for asymmetry in associations, as *p(R|C)* reflects the extent to which the topics in which the cue appears give high probability to the response, and high-frequency words tend to appear in more topics than low-frequency words. In other words, we can obtain *p(baby|stork)* > *p(stork|baby)* as *baby* is a high frequency word with many associated topics, whereas *stork* is a low frequency word with few associated topics.

In fact, free association derived from LDA performs well in quantitative evaluation on free association norms. Nematzadeh et al. (2017) found that LDA trained on the smallest of their three corpora was comparable to Word2Vec-SkipGram or GloVe trained on the same corpus, for coefficient of correlation (LDA r = .2, Word2Vec r = .25, GloVE r = .2) and median rank of associates (LDA 23/69/103.5, Word2Vec 26/72/106, GloVe 56/138/215), but much better than Word2Vec and GloVe on asymmetry ratio (LDA r = .49 vs GloVe r = .32 and Word2Vec r = .01). Of course, Word2Vec and especially GloVe achieved better performance than this when trained on larger corpora (except asymmetry ratio, where LDA still beat out the best GloVe model's *r* = .48). Unfortunately, Nematzadeh et al. found it difficult to train an effective LDA model on these larger corpora. They found that Markov chain Monte Carlo methods (in particular, Gibbs sampling) for fitting the LDA model did not scale to larger corpora, whereas more efficient variational inference methods applied to larger corpora produced models of poor quality. In our tests, we had similar difficulties, even when using MCMC to fit LDA to smaller corpora with off-the-shelf tools (including when using more efficient Hamiltonian Monte Carlo as implemented in the *pymc3* package; Salvatier et al., 2016). Therefore, we compared our GloVe-based models to LDA fit with variational inference, but we advise caution in interpreting these results in light of these difficulties.

Implementation Details

We fit an LDA model to the Touchstone Applied Science Associates (TASA) corpus (Landauer & Dumais, 1997), which is a collection of passages excerpted from educational texts used in curricula from the first year of school to the first year of college. This corpus contains approximately 8 million words across nearly 40,000 documents. This is the same corpus to which Griffiths et al. (2007) and Nematzadeh et al. (2017) fit LDA for modeling free association. Following their pre-processing steps, we considered only words that (a) occurred at least 10 times in TASA and (b) were not included in a standard list of stop-words containing function words and other high-frequency words with little semantic content. This left us with about 25,000 unique words in the TASA corpus.

We fit LDA to the TASA corpus with the online variational inference algorithm implemented in scikit-learn (Pedregosa et al., 2011). Following Griffiths et al. (2007), we used 1,700 topics (we also tried 10% as many topics and results were generally the same). All other hyperparameters were left to their defaults. Unfortunately, many of the cue or response words from our first-response test set were missing from this model (either because they didn't meet our frequency threshold, or, less likely, because they were stop words), and so the resulting LDA model could only make predictions for 7,520 of the 10,035 trials in the test set.

Results

Results are in Table S1. For ease of comparison, we have reproduced our primary GloVe-based models' results on the test set we constructed from SWOW-EN. Crucially, because the number of evaluable trials in the test set differs between our LDA model (7,520) and our GloVe-based models (10,035), we report trial-averaged and median negative log likelihood, since this controls for the different numbers of test trials. Otherwise, calculation of performance metrics is identical to that in the main text.

Table S1. Performance of both first response model variants and LDA-VI on all model evaluation metrics, on our test set selected from SWOW-EN. 95% confidence intervals are in parentheses.

Evaluation Metric		**Asymmetric weights**	**Symmetric weights**	**LDA-VI**
Average trial-level negative log-likelihood		8.64	11.48	7.84
Median trial-level negative log-likelihood		7.21	10.56	8.23
$p(R_1\|C)$ correlation		.37 (.37, .40)	.29 (.26, .33)	.06 (.02, .10)
Median rank of true associates in model predictions	1st associate	5	23	441
	2nd associate	11	31	657
	3rd associate	46	86	723
Asymmetry ratio correlation		.59 (.58, .61)	.38 (.36, .40)	.08 (-.03, .19)
Asymmetry direction accuracy (in %)		79 (78, 80)	67 (66, 68)	3 (1, 5)

As can be seen, LDA generally performs very poorly, except, curiously, on average negative log likelihood, where it is better than both GloVe-based models. One possible explanation for this is that the LDA model had a smaller number of possible responses (3,092) compared to the GloVe-based models (3,616), and was evaluated the LDA model on a narrower set of test trials only limited to its possible responses. This means that the LDA model can, in principle, achieve better negative log likelihood since random choices are more likely to be correct. There is also another explanation. In an analysis of our models' predictions, we found that our CMR-based models tended to concentrate more probability on their top predicted response, on average (across cues) putting about .6 or .7 (depending on symmetric or asymmetric weights) on the response that these models thought was most likely. The average probability for the 2nd, 3rd, etc. responses dropped off to ~.1 and below. The LDA model, however, made much more diffuse predictions, putting on average only .05 probability on its most probable response, and subsequent ranks dropped off more gradually. Essentially, this meant that, on the trials in which our model made incorrect predictions, it was penalized quite heavily in terms of log likelihood. These trials were fairly infrequent, and consisted mostly of atypical or non-modal cue-response combinations (i.e. responses generated by idiosyncratic participants, that differed greatly from the modal population response to the cue). However, through averaging, these trials led to a substantial reduction in the performance of the CMR models relative to LDA. Consistent with this explanation, we found that when examining median trial-level log likelihoods, the performance of our asymmetric weights model exceeded that of LDA. This is also why all other metrics (which are at the level of unique cues or unique cue-response pairs in the test set) favor our CMR models over LDA.

Overall, we suspect that the generally poor results of LDA are due to fitting with variational inference instead of MCMC. As we cited earlier, Nematazadeh et al. (2017) reported that LDA fit with variational inference performed poorly on free association tests, so it is not surprising that we find the same result here. Therefore, we cannot claim that our CMR models are superior to LDA-based models in particular or topic models more generally in explaining free association.

References

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

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review, 114*(2), 211.

Nematzadeh, A., Meylan, S. C., & Griffiths, T. L. (2017). Evaluating vector-space models of word representation, or, the unreasonable effectiveness of counting words near other words. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review, 104*(2), 211.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Salvatier J., Wiecki T.V., Fonnesbeck C. (2016) Probabilistic programming in Python using PyMC3. *PeerJ Computer Science* 2:e55. DOI: 10.7717/peerj-cs.55.