Pulling Out the Stops: Rethinking Stopword Removal for Topic Models

Alexandra Schofield
Cornell University
Ithaca, NY 14850
xanda@cs.cornell.edu

Måns Magnusson
Linköping University
Linköping, Sweden
mans.magnusson@liu.se

David Mimno
Cornell University
Ithaca, NY 14850
mimno@cornell.edu

Abstract

It is often assumed that topic models benefit from the use of a manually curated stopword list. Constructing this list is time-consuming and often subject to user judgments about what kinds of words are important to the model and the application. Although stopword removal clearly affects which word types appear as most probable terms in topics, we argue that this improvement is superficial, and that topic inference benefits little from the practice of removing stopwords beyond very frequent terms. Removing corpus-specific stopwords after model inference is more transparent and produces similar results to removing those words prior to inference.

1 Introduction

In Latent Dirichlet allocation (LDA) (Blei et al., 2003), a common preprocessing step is the removal of stopwords, or common, contentless words in a corpus. The use of stoplists comes with several costs in both effort and persuasiveness. Constructing a good stoplist is difficult and time consuming, and often cannot be transferred to new corpora. Custom stoplists can also call into question the validity of a model: if an analyst is too aggressive in removing words, the resulting models may be biased towards what the analyst views as important in a corpus. Finally, while removing stopwords appears to produce more interpretable topics, this effect may be an illusion. As topic interpretability is typically judged by the most frequent terms in the topic, post-hoc stopword removal from a model can substantially increase interpretability without modifying the model.

In this paper, we analyze the consequence of removing stopwords for topic modeling in terms of model fit, coherence, and utility. We consider three configurations: models trained and presented with stopwords intact, models with stopwords removed *before* training, and models with stopwords removed *after* training. We find that there are benefits in model quality when stopwords are removed. However, stopword removal does not appear to consistently improve the model's ability to learn topics over the other terms, but rather to remove dense high-probability terms that can slow inference and skew the word type probability distribution. We conclude that beyond high-probability terms, the effects of stoplists on training are limited, and that removing unwanted terms after training should be sufficient.

2 Stopwords in Topic Models

The assumption behind stopword removal is that, with stopwords present, we will not be able to learn as high-quality a language model. In the corpora we have selected, a preset list of approximately 500 stopword types accounted for 40-50% of the corpus. If these words are expected to be uncorrelated with any topics, we would expect stopwords to only hinder inference of meaningful topics. LDA may sometimes partially accommodate separating out stopwords without explicitly removing them. Wallach et al. (2009a) show that a parsimonious asymmetric Dirichlet prior inferred for θ, can allow model inference to isolate stopwords into fewer low-quality topics, leaving the remaining topics largely unaffected.

In essence, these low-quality topics learn a background distribution for stopwords, but infrequent contentless words may be inadvertently correlated with contentful topic terms, while words such as "the" are so frequent they are still likely to be prominent in many topics. The former terms, by virtue of being infrequent, should not disrupt

topics, but the latter set of extremely frequent terms may overwhelm the model and reduce how well the model fits contentful terms.

We identify three plausible hypotheses about the effect of stopwords in topic model training.

1. **Stopwords harm inference.** Noise from frequent words prevents the algorithm from recognizing patterns in content-bearing words.

2. **Stopwords have no effect on inference.** Noise from frequent words does not alter inference on non-stopwords.

3. **Stopwords improve inference.** Frequent words echo and reinforce patterns in content-bearing words.

We assess through a variety of experiments how well each of these hypotheses hold in practice.

3 Evaluation Methods

We aim to study the effects of removing stopwords on topic quality and keyword generation. To do this, we evaluate topic models as language models, document summarization tools, and features for learning new models over data.

3.1 Existing Methods

A standard measurement of topic model quality is based upon evaluating the likelihood of a held-out portion of the modeled corpus being generated by the inferred topic model (Wallach et al., 2009b). Though directly computing a document's probability in an LDA model is intractable, we can estimate it using left-to-right estimation (Wallach et al., 2009b). However, this metric has two drawbacks: one, that it provides little information about individual topics, and two, that it does not correlate well with actual human perception of topic quality (Chang et al., 2009).

Work demonstrating that topic likelihood and human evaluations of topic coherence differ (Chang et al., 2009) has led to several metrics to evaluate a topic's coherence. These typically use co-occurrence statistics for frequent types in the topic, such as topic coherence (Mimno et al., 2011) and normalized pointwise mutual information (NPMI) (Aletras and Stevenson, 2013; Lau et al., 2014). We use NPMI in our evaluations.

3.2 New Methods

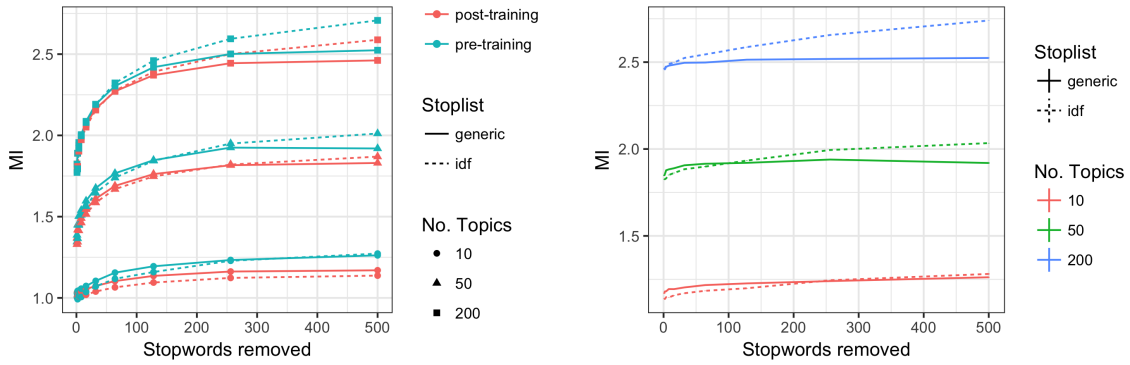
Topic-document mutual information The hypotheses described in Section 2 focus on differences between the topic distribution of stopwords in a given document and the topic distribution of content-bearing words in that document. One way to assess this effect in a model is to study the mutual information between documents and topics. Using the topic assignments of tokens inferred via Gibbs sampling, we can examine the mutual information between the document-topic distribution and the topic assignment of the token. We compare the $MI(d, k)$ before and after stopword removal to measure the effect of removal on the posterior. If there is no semantic information in a set of tokens (such as stopwords) the $MI(d, k)$ should be close to 0. If the stopwords have a negative effect on inference (hypothesis 1) removing these words before inference (*pre*-removal) should result in a higher $MI(d, k)$ than removing them afterwards (*post*-removal). The opposite should be true if stopword improve inference (hypothesis 3).

Classification with key terms A metric of the quality of representative terms for a topic is their ability to identify documents with a high proportion of that topic. Inspired by the approach of Dredze et al. (2008), we use classification of documents by topic to assess the quality of key terms as representative topic features. We train multinomial Naïve Bayes models with the token counts of top representative terms as features and the most present topic of each document as labels.

4 Experiments

We evaluate the results of removing stopwords for topic modeling on two different corpora: a corpus of United States State of the Union (SOTU) addresses from 1790 to 2009 split into paragraphs, and a 1% sample of the New York Times Annotated corpus (Sandhaus, 2008), spanning articles from 1987 to 2007 and split into 500-word segments to handle overly-long articles. For experiments relying on held-out data for the NYT corpus, we sampled approximately 5% of the articles to be used as a testing corpus. We treat the full article set as a reference corpus for word co-occurrence. The details of the size of each corpus are in Table 1. We use a standard stoplist from MALLET for our experiments (McCallum, 2002).

Our experiments use topic models trained with



(a) Pre- and post removal

(b) Effect on non-stopword tokens

Figure 1: Mutual information on the NYT corpus, $MI(d, k)$, as a function of the number of stopwords removed (ordered by number of tokens). Removing stopwords before training leads to a slightly higher MI, but the effect on non-stopword tokens is small.

MALLET (McCallum, 2002). We inferred LDA topic models of 10, 50, and 200 topics with 1000 iterations of Gibbs sampling using hyperparameter optimization (Wallach et al., 2009a). Topics were trained on versions of the corpora with and without stopwords, with an additional model inferred by recomputing the document-topic distributions θ and topic-word distributions ϕ after removing all stopwords from the inferred topic assignments of models trained with stopwords. This allows us to compare models with no stopwords removed (control), stopwords removed before training (pre), and stopwords removed after training (post) all with the same effective corpus. Metrics are averaged over 10 models per treatment.

We train several types of topic models on New York Times (NYT) data. Our standard treatment defines a document as one full article, but we additionally train models on a segmented version of the corpus (NYT-S) where each article is broken into 100-word segments. In addition, we include models with unoptimized hyperparameters (NYT-U), set as $\sum_k \alpha_k = 5$ and $\beta = 0.01$.

Corpus	Documents	Tokens
NYT	18820	10.33M
NYT-S	18820	6.50M
SOTU	19254	1.264M
SOTU-S	19254	681K

Table 1: Details of the New York Times (NYT) and State of the Union (SOTU) corpora used for topic modeling. We experiment a fixed English stoplist of 524 words to remove stopwords (-S). We use the full SOTU corpus for training.

4.1 Mutual Information

In Figure 1, we examine topic-document mutual information for different sized sets of stopwords removed before and after training the model. By removing stopwords before training, we obtain a slightly higher $MI(d, k)$ than removing stopwords after training, in support of hypothesis 1 in Section 2. However, this difference is relatively small compared to including more stopwords or changing the number of topics.

If we focus on terms besides stopwords, we can see that the effect of removing stopwords is relatively small. There is some difference in removing the most common stopwords, but extending a stoplist has diminishing returns, supporting hypothesis 2 in Section 2.

4.2 Log Likelihood

In order to better evaluate the effect of stopword removal on improving model training, we compare the inferred log-likelihood of models trained on our 1% New York Times sample on our larger 5% testing sample. As seen in Table 2, the choice of when to remove stopwords has little effect. On

Topics	pre	post
10	-10.830 ± 0.006	-10.826 ± 0.005
50	-10.708 ± 0.007	-10.702 ± 0.007
200	-10.532 ± 0.002	-10.529 ± 0.002

Table 2: Per-token log likelihood measures on held-out data for New York Times models with standard error. Removing stopwords before training (pre) does not statistically significantly differ from removing stopwords after training (post).

pre	1 num art museum work show artists works artist paintings exhibition gallery painting arts american collection 2 num beloved paid family notice wife deaths husband late loving memorial funeral devoted service services 3 life love world story sense young man makes good style full real beautiful dark turns
post	1 num art museum show artists work works exhibition gallery artist paintings arts painting american collection 2 family president board passing love friend paid member notice jewish beloved chairman miss condolences deaths 3 book life story man books young love written world characters character work writing james author

Table 3: Example topics from 50-topic New York Times models with stopwords removed before and after training. Post-removal topics look similar but lack some more common terms found with pre-removal.

Topics	Treatment	control	pre	post
10	NYT	0.0280	0.0874	0.0931
	NYT-S	0.0282	0.0850	0.0851
	NYT-U	0.0311	0.0863	0.0878
	SOTU	0.0248	0.0406	0.0402
50	NYT	0.0595	0.1271	0.1209
	NYT-S	0.0531	0.1257	0.1195
	NYT-U	0.0554	0.1233	0.1208
	SOTU	0.0438	0.0655	0.0612
200	NYT	0.0951	0.1352	0.1317
	NYT-S	0.0718	0.1317	0.1239
	NYT-U	0.1021	0.1352	0.1338
	SOTU	0.0542	0.0681	0.0637

Table 4: The average NPMI scores for New York Times and State of the Union data. Surprisingly, with 10 topics, post-removal of stopwords often produces better coherence.

the New York Times held-out dataset, the effect of post-removing stopwords after training is statistically indistinguishable from pre-removing them. This supports hypothesis 2 in Section 2, that stopwords are not actually significantly affecting the model inference process for other terms.

4.3 Coherence

We report the average NPMI scores for the New York Times and State of the Union data in Table 4. While removing stopwords from the top keys for coherence evaluation improves model coherence over the control, again, the choice of when the stopwords are removed from the vocabulary seems to have very little effect. Especially for only 10 topics, coherence of models where stopwords were removed after training can slightly outperform models with pre-removal. This finding supports hypothesis 2 over hypothesis 1 in Section 2: though removal of stopwords before training improves automatically-evaluated coherence, *when* they are removed has little impact.

4.4 Classification with Key Terms

We use the 15 most probable words from each 50-topic model on New York Times sample data to train a logistic regression classifier to recognize the most prominent topic for each document. We use 10-fold cross validation to compute accuracy, which we report in Table 5. Unsurprisingly, removing stopwords at some stage improves the classification accuracy of key terms. However, we note that removing terms before training is significantly better only for one of the four treatments (NYT-S) and is actually significantly worse than removing after for the standard NYT setting. This again supports hypothesis 2 in Section 2: removing the stopwords before training does not alter the distinctiveness of topics based on high-probability terms.

	control	pre	post
NYT	$47.1 \pm 0.3\%$	$69.4 \pm 0.2\%$	$69.9 \pm 0.2\%$
NYT-S	$47.1 \pm 0.2\%$	$54.0 \pm 0.2\%$	$53.3 \pm 0.1\%$
NYT-U	$62.6 \pm 0.3\%$	$69.8 \pm 0.2\%$	$69.8 \pm 0.2\%$
SOTU	$43.8 \pm 0.3\%$	$48.7 \pm 0.2\%$	$48.8 \pm 0.2\%$

Table 5: Classification results using top terms of 50-topic models on NYT and SOTU data. Removing stopwords is often equally effective before and after training.

Examples of topics in Table 3 provide some depth to understanding these results. Topic 1 is nearly identical across the two treatments, while topic 3 uses terms clearly from reviews when stopwords are removed before that seem to be lost when stopwords are removed afterwards. Anecdotally, common content words appear not to be modeled as well when stopwords are present.

5 Conclusion

Our results demonstrate that, as per our second hypothesis, removing stopwords *after* training is generally just as effective as removing them before. Rather than leading the model to infer more coherent topics by removing words that we expect to have no content, removing stopwords appears to simply reduce the amount of probability mass and smoothing of the model caused by frequent non-topic-specific terms.

Consequently, generating a corpus-specific

stoplist to remove rarer contentless words provides relatively little utility to training a model. To obtain the benefit of a stoplist, it suffices to remove the most frequent, obvious stopwords from a corpus without developing a specific stoplist for the problem setting. If these methods are not sufficient, we find that post-hoc stopword removal can significantly improve coherence while avoiding many of the efficiency and epistemological bias issues of iterative stoplist curation. We believe this result will be beneficial for researchers in other fields navigating the pragmatics of using topic models for their own investigations.

6 Acknowledgments

This work was supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program and a fellowship from the Alfred P. Sloan Foundation.

References

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 13–22, Potsdam, Germany, March. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, March.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 288–296. Curran Associates, Inc.

Mark Dredze, Hanna M. Wallach, Danny Puller, and Fernando Pereira. 2008. Generating summary keywords for emails using topics. In *Proceedings of the 13th International Conference on Intelligent User Interfaces*, IUI '08, pages 199–206, New York, NY, USA. ACM.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden, April. Association for Computational Linguistics.

Andrew K. McCallum. 2002. MALLET: a machine learning for language toolkit. Available at: http://mallet.cs.umass.edu.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Evan Sandhaus. 2008. The New York Times annotated corpus. *Linguistic Data Consortium*, DVD: LDC2009T19.

Hanna M. Wallach, David M. Mimno, and Andrew McCallum. 2009a. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. Curran Associates, Inc.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009b. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1105–1112, New York, NY, USA. ACM.