Coherence-Aware Neural Topic Modeling

Ran Ding, Ramesh Nallapati, Bing Xiang
Amazon Web Services
{rding, rnallapa, bxiang}@amazon.com

Abstract

Topic models are evaluated based on their ability to describe documents well (i.e. low perplexity) and to produce topics that carry coherent semantic meaning. In topic modeling so far, perplexity is a direct optimization target. However, topic coherence, owing to its challenging computation, is not optimized for and is only evaluated after training. In this work, under a neural variational inference framework, we propose methods to incorporate a topic coherence objective into the training process. We demonstrate that such a coherence-aware topic model exhibits a similar level of perplexity as baseline models but achieves substantially higher topic coherence.

1 Introduction

In the setting of a topic model (Blei, 2012), perplexity measures the model's capability to describe documents according to a generative process based on the learned set of topics. In addition to describing documents well (i.e. achieving low perplexity), it is desirable to have topics (represented by top-N most probable words) that are human-interpretable. Topic interpretability or coherence can be measured by *normalized point-wise mutual information* (NPMI) (Aletras and Stevenson, 2013; Lau et al., 2014). The calculation of NPMI however is based on look-up operations in a large reference corpus and therefore is non-differentiable and computationally intensive. Likely due to these reasons, topic models so far have been solely optimizing for perplexity, and topic coherence is only evaluated after training. As has been noted in several publications (Chang et al., 2009), optimization for perplexity alone tends to negatively impact topic coherence. Thus, without introducing topic coherence as a training objective, topic modeling likely produces sub-optimal results.

Compared to classical methods, such as mean-field approximation (Hoffman et al., 2010) and collapsed Gibbs sampling (Griffiths and Steyvers, 2004) for the latent Dirichlet allocation (LDA) (Blei et al., 2003) model, neural variational inference (Kingma and Welling, 2013; Rezende et al., 2014) offers a flexible framework to accommodate more expressive topic models. We build upon the line of work on topic modeling using neural variational inference (Miao et al., 2016, 2017; Srivastava and Sutton, 2017) and incorporate topic coherence awareness into topic modeling.

Our approaches of constructing topic coherence training objective leverage pre-trained word embeddings (Mikolov et al., 2013; Pennington et al., 2014; Salle et al., 2016; Joulin et al., 2016). The main motivation is that word embeddings carry contextual similarity information that is highly related to the mutual information terms involved in the calculation of NPMI. In this paper, we explore two methods: (1) we explicitly construct a differentiable surrogate topic coherence regularization term; (2) we use word embedding matrix as a factorization constraint on the topical word distribution matrix that implicitly encourages topic coherence.

2 Models

2.1 Baseline: Neural Topic Model (NTM)

The model architecture shown in Figure 1 is a variant of the Neural Variational Document Model (NVDM) (Miao et al., 2016). Let $x \in \mathbb{R}^{|V| \times 1}$ be the bag-of-words (BOW) representation of a document, where $|V|$ is the size of the vocabulary and let $z \in \mathbb{R}^{K \times 1}$ be the latent topic variable, where K is the number of topics. In the encoder $q_\phi(z|x)$, we have $\pi = f_{MLP}(x)$, $\mu(x) = l_1(\pi)$, $\log \sigma(x) = l_2(\pi)$, $h(x, \epsilon) = \mu + \sigma \odot \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$, and finally $z = f(h) = \text{ReLU}(h)$. The functions l_1 and l_2 are linear transformations

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 830–836
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics

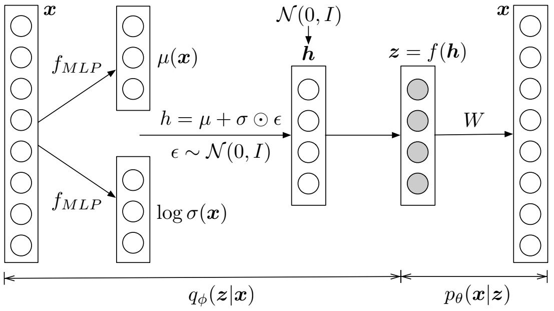

Figure 1: Model architecture

with bias. We choose the multi-layer perceptron (MLP) in the encoder to have two hidden layers with $3 \times K$ and $2 \times K$ hidden units respectively, and we use the sigmoid activation function. The decoder network $p_\theta(x|z)$ first maps z to the predicted probability of each of the word in the vocabulary $y \in \mathbb{R}^{|V| \times 1}$ through $y = \text{softmax}(Wz + b)$, where $W \in \mathbb{R}^{|V| \times K}$. The log-likelihood of the document can be written as $\log p_\theta(x|z) = \sum_{i=1}^{|V|} \{x \odot \log y\}$. We name this model Neural Topic Model (NTM) and use it as our baseline. We use the same encoder MLP configuration for our NVDM implementation and all variants of NTM models used in Section 3. In NTM, the objective function to maximize is the usual *evidence lower bound* (ELBO) which can be expressed as

$$\mathcal{L}_{ELBO}(x^i) \approx \frac{1}{L} \sum_{l=1}^{L} \log p_\theta(x^i | z^{i,l}) - D_{KL}(q_\phi(h|x) || p_\theta(h))$$

where $z^{i,l} = \text{ReLU}(h(x^i, \epsilon^l))$, $\epsilon^l \sim \mathcal{N}(0, I)$. We approximate $\mathbb{E}_{z \sim q(z|x)}[\log p_\theta(x|z)]$ with Monte Carlo integration and calculate the Kullback-Liebler (KL) divergence analytically using the fact $D_{KL}(q_\phi(z|x)||p_\theta(z)) = D_{KL}(q_\phi(h|x)||p_\theta(h))$ due to the invariance of KL divergence under deterministic mapping between h and z.

Compared to NTM, NVDM uses different activation functions and has $z = h$. Miao (2017) proposed a modification to NVDM called Gaussian Softmax Model (GSM) corresponding to having $z = \text{softmax}(h)$. Srivastava (2017) proposed a model called ProdLDA, which uses a Dirichlet prior instead of Gaussian prior for the latent variable h. Given a learned W, the practice to extract top-N most probable words for each topic is to take the most positive entries in each column of W (Miao et al., 2016, 2017; Srivastava and Sutton, 2017). This is an intuitive choice, provided that z is non-negative, which is indeed the case for NTM, GSM and ProdLDA. NVDM, GSM, and ProdLDA are state-of-the-art neural topic models which we will use for comparison in Section 3.

2.2 Topic Coherence Regularization: NTM-R

The topic coherence metric NPMI (Aletras and Stevenson, 2013; Lau et al., 2014) is defined as

$$\text{NPMI}(\boldsymbol{w}) = \frac{1}{N(N-1)} \sum_{j=2}^{N} \sum_{i=1}^{j-1} \frac{\log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)}$$

where $\boldsymbol{w}$ is the list of top-N words for a topic. N is usually set to 10. For a model generating K topics, the overall NPMI score is an average over all topics. The computational overhead and non-differentiability originate from extracting the co-occurrence frequency from a large corpus[1].

From the NPMI formula, it is clear that word-pairs that co-occur often would score high, unless they are rare word-pairs – which would be normalized out by the denominator. The NPMI scoring bears remarkable resemblance to the contextual similarity produced by the inner product of word embedding vectors. Along this line of reasoning, we construct a differentiable, computation-efficient word embedding based topic coherence (WETC).

Let E be the row-normalized word embedding matrix for a list of N words, such that $E \in \mathbb{R}^{N \times D}$ and $\|E_{i,:}\| = 1$, where D is the dimension of the embedding space. We can define *pair-wise* word embedding topic coherence in a similar spirit as NPMI:

$$\text{WETC}_{PW}(E) = \frac{1}{N(N-1)} \sum_{j=2}^{N} \sum_{i=1}^{j-1} \langle E_{i,:}, E_{j,:} \rangle = \frac{\sum\{E^T E\} - N}{2N(N-1)}$$

where $\langle \cdot, \cdot \rangle$ denotes inner product. Alternatively, we can define *centroid* word embedding topic coherence

$$\text{WETC}_C(E) = \frac{1}{N} \sum \{E t^T\}$$

where vector $t \in \mathbb{R}^{1 \times D}$ is the centroid of E, normalized to have $\|t\| = 1$. Empirically, we found

[1] A typical calculation of NPMI over 50 topics based on the Wikipedia corpus takes ∼20 minutes, using code provided by (Lau et al., 2014) at https://github.com/jhlau/topic_interpretability.

that the two WETC formulations behave very similarly. In addition, both WETC$_{PW}$ and WETC$_C$ correlate to human judgement almost equally well as NPMI when using `GloVe` (Pennington et al., 2014) vectors[2].

With the above observations, we propose the following procedure to construct a WETC-based surrogate topic coherence regularization term: (1) let $E \in \mathbb{R}^{|V| \times D}$ be the pre-trained word embedding matrix for the vocabulary, rows aligned with W; (2) form the W-weighted centroid (topic) vectors $T \in \mathbb{R}^{D \times K}$ by $T = E^T W$; (3) calculate the cosine similarity matrix $S \in \mathbb{R}^{|V| \times K}$ between word vectors and topic vectors by $S = ET$; (4) calculate the W-weighted sum of word-to-topic cosine similarities for each topic $C \in \mathbb{R}^{1 \times K}$ as $C = \sum_i (S \odot W)_{i,:}$. Compared to WETC$_C$, in calculating C we do not perform top-N operation in W, but directly use W for weighted sum. Specifically, we use W-weighted topic vector construction in Step-2 and W-weighted sum of the cosine similarities between word vectors and topic vectors in Step-4. To avoid unbounded optimization, we normalize the rows of E and the columns of W before Step-2, and normalize the columns of T after Step-2. The overall maximization objective function becomes $\mathcal{L}_R(x; \theta, \phi) = \mathcal{L}_{ELBO} + \lambda \sum_i C_i$, where λ is a hyper-parameter with positive values controlling the strength of topic coherence regularization. We name this model NTM-R.

2.3 Word Embedding as a Factorization Constraint: NTM-F and NTM-FR

Instead of allowing all the elements in W to be freely optimized, we can impose a factorization constraint of $W = E\hat{T}$, where E is the pre-trained word embedding matrix that is *fixed*, and only $\hat{T}$ is allowed to be learned through training. Under this configuration, $\hat{T}$ lives in the embedding space, and each entry in W is the dot product similarity between a topic vector $\hat{T}_i$ and a word vector E_j. As one can imagine, similar words would have similar vector representations in E and would have similar weights in each column of W. Therefore the factorization constraint encourages words with similar meaning to be selected or de-selected together thus potentially improving topic coherence.

We name the NTM model with factorization constraint enabled as NTM-F. In addition, we can apply the regularization discussed in the previous section on the resulting matrix W and we name the resulting model NTM-FR.

Metric	Perplexity		NPMI	
Number of topics	50	200	50	200
LDA				
LDA, mean-field	1046	1195	0.11	0.06
LDA, collapsed Gibbs	**728**	**688**	0.17	0.14
Neural Models				
NVDM	750	743	0.14	0.13
GSM	787	829	0.22	0.19
ProdLDA	1172	1168	0.28	0.24
NTM	780	768	0.18	0.18
NTM-R	**775**	**763**	**0.28**	**0.23**
NTM-F	898	1086	**0.29**	0.24
NTM-FR	924	1225	0.27	**0.26**

Table 1: Comparison to LDA and neural variational models on the *20NewsGroup* dataset. Best numbers are bolded. The blue underlined row highlights the best NPMI and perplexity tradeoff as discussed in text.

3 Experiments and Discussions

3.1 Results on *20NewsGroup*

First, we compare the proposed models to state-of-the-art neural variational inference based topic models in the literature (NVDM, GSM, and ProdLDA) as well as LDA benchmarks, on the *20NewsGroup* dataset[3]. In training NVDM and all NTM models, we used `Adadelta` optimizer (Zeiler, 2012). We set the learning rate to 0.01 and train with a batch size of 256. For NTM-R, NTM-F and NTM-FR, the word embedding we used is `GloVe` (Pennington et al., 2014) vectors pre-trained on Wikipedia and Gigaword with 400,000 vocabulary size and 50 embedding dimensions[4]. The topic coherence regularization coefficient λ is set to 50. The results are presented in Table 1.

Overall we can see that LDA trained with collapsed Gibbs sampling achieves the best perplexity, while NTM-F and NTM-FR models achieve the best topic coherence (in NPMI). Clearly, there is a trade-off between perplexity and NPMI as identified by other papers. So we constructed Figure 2, which shows the two metrics from various models. For the models we implemented, we additionally

[2] See Appendix A for details on an empirical study of human judgement of topic coherence, NPMI and WETC with various types of word embeddings.

[3] We use the exact dataset from (Srivastava and Sutton, 2017) to avoid subtle differences in pre-processing

[4] Obtained from `https://nlp.stanford.edu/projects/glove/`

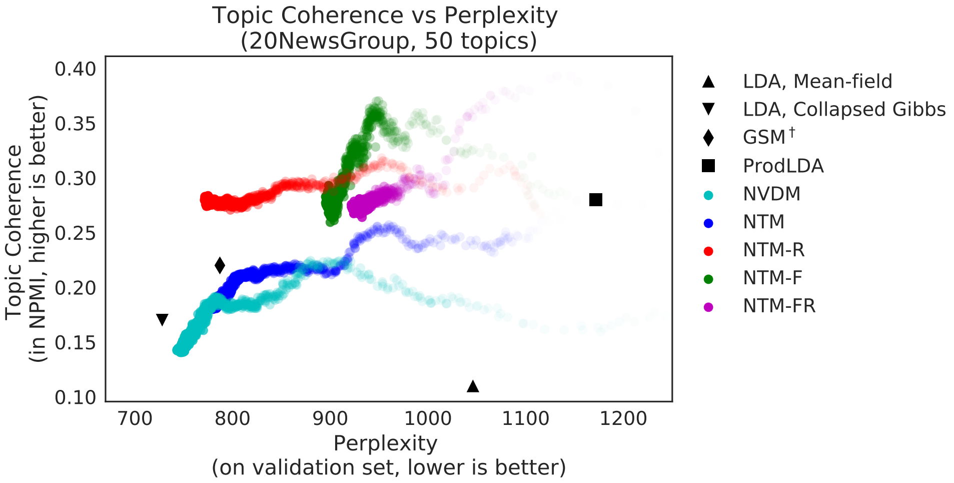

Figure 2: NPMI vs. perplexity for various models at 50 topics. For NVDM and NTM models the traces correspond to the evolution over training epochs. High transparency is the beginning of the training.

show the full evolution of these two metrics over training epochs.

From Figure 2, it becomes clear that although ProdLDA exhibits good performance on NPMI, it is achieved at a steep cost of perplexity, while NTM-R achieves similar or better NPMI at much lower perplexity levels. At the other end of the spectrum, if we look for low perplexity, the best numbers among neural variational models are between 750 and 800. In this neighborhood, NTM-R substantially outperforms the GSM, NVDM and NTM baseline models. Therefore, we consider NTM-R the best model overall. Different downstream applications may require different tradeoff points between NPMI and perplexity. However, the proposed NTM-R model does appear to provide tradeoff points on a Pareto front compared to other models across most of the range of perplexity.

3.2 Comments on NTM-F and NTM-FR

It is worth noting that although NTM-F and NTM-FR exhibit high NPMI early on, they fail to maintain it during the training process. In addition, both models converged to fairly high perplexity levels. Our hypothesis is that this is caused by NTM-F and NTM-FR's substantially reduced parameter space - from $|V| \times K$ to $D \times K$, where $|V|$ ranges from 1,000 to 150,000 in a typical dataset, while D is on the order of 100.

Some form of relaxation could alleviate this problem. For example, we can let $W = E\hat{T} + A$, where A is of size $|V| \times K$ but is heavily regularized, or let $W = EQ\hat{T}$ where Q is allowed as additional free parameters. We leave fully addressing this to future work.

3.3 Validation on other Datasets

To further validate the performance improvement from using WETC-based regularization in NTM-R, we compare NTM-R with the NTM baseline model on a few more datasets: DailyKOS, NIPS, and NYTimes[5] (Asuncion and Newman, 2007). These datasets offer a wide range of document length (ranging from ∼100 to ∼1000 words), vocabulary size (ranging from ∼7,000 to ∼140,000), and type of documents (from news articles to long-form scientific writing). In this set of experiments, we used the same settings and hyperparameter λ as before and did not fine-tune for each dataset. The results are presented in Figure 3. In a similar style as Figure 2, we show the evolution of NPMI and WETC versus perplexity over epochs until convergence.

Among all datasets, we observed improved NPMI at the same perplexity level, validating the effectiveness of the topic coherence regularization. However, on the NYTimes dataset, the improvement is quite marginal even though WETC improvements are very noticeable. One particularity about the NYTimes dataset is that approximately 58,000 words in the 140,000-word vocabulary are named entities. It appears that the large number of named entities resulted in a divergence between NPMI and WETC scoring, which is an issue to address in the future.

[5] https://archive.ics.uci.edu/ml/datasets/Bag+of+Words

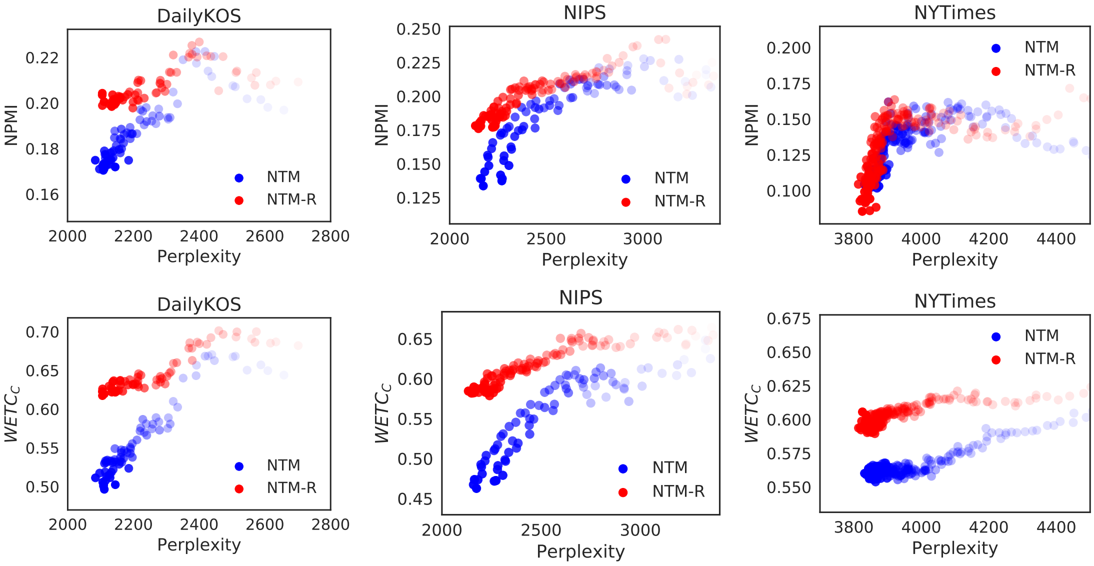

Figure 3: Performance comparison between NTM-R and NTM on multiple datasets, with 50 topics. Top row is NPMI versus perplexity, bottom row is WETC_C versus perplexity. From left to right: DailyKOS, NIPS, and NYTimes. See text for details about the datasets.

4 Conclusions

In this work, we proposed regularization and factorization constraints based approaches to incorporate awareness of topic coherence into the formulation of topic models: NTM-R and NTM-F respectively. We observed that NTM-R substantially improves topic coherence with minimal sacrifice in perplexity. To our best knowledge, NTM-R is the first topic model that is trained with an objective towards topic coherence – a feature directly contributing to its superior performance. We further showed that the proposed WETC-based regularization method is applicable to a wide range of text datasets.

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.

Arthur Asuncion and David Newman. 2007. UCI Machine Learning Repository.

David M Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, pages 288–296.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.

Matthew Hoffman, Francis R Bach, and David M Blei. 2010. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 856–864.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

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.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *International Conference on Machine Learning*, pages 2410–2419.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *International Conference on Machine Learning*, pages 1727–1736.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286.

Alexandre Salle, Marco Idiart, and Aline Villavicencio. 2016. Matrix factorization using window sampling and negative sampling for improved word representations. *arXiv preprint arXiv:1606.00819*.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

A Word Embedding Topic Coherence

As studied in (Aletras and Stevenson, 2013) and (Lau et al., 2014), the NPMI metric for assessing topic coherence over a list of words $\boldsymbol{w}$ is defined in Eq. 1.

$$\mathrm{NPMI}(\boldsymbol{w}) = \frac{1}{N(N-1)}\sum_{j=2}^{N}\sum_{i=1}^{j-1}\frac{\log\frac{P(w_i,w_j)}{P(w_i)P(w_j)}}{-\log P(w_i,w_j)} \quad (1)$$

where $P(w_i)$ and $P(w_i, w_j)$ are the probability of words and word pairs, calculated based on a reference corpus. N is usually set to 10, by convention, so that NPMI is evaluated over the topic-10 words for each topic. For a model generating K topics, the overall NPMI score is an average over all the topics. The computational overhead comes from extracting the relevant co-occurrence frequency from a large corpus. This problem is exacerbated when the look-up also requires a small sliding window as the authors of (Lau et al., 2014) suggested. A typical calculation of 50 topics based on a few million documents from the Wikipedia corpus takes $\sim$20 minutes[6].

For a list of words $\boldsymbol{w}$ of length N, we can assemble a corresponding word embedding matrix $E \in \mathbb{R}^{N\times D}$ with each row corresponding to a word in the list. D is the dimension of the embedding space. Averaging across the rows, we can obtain vector $t \in \mathbb{R}^{1\times D}$ as the centroid of all the word vectors. It may be regarded as a "topic" vector. In addition, we assume that each row of E and t is normalized, i.e. $\|t\| = 1$ and $\|E_{i,:}\| = 1$. With these, we define *pair-wise* and *centroid* word embedding topic coherence WETC_{PW} and WETC_C as follows:

$$\begin{aligned}\mathrm{WETC}_{PW}(E) &= \frac{1}{N(N-1)}\sum_{j=2}^{N}\sum_{i=1}^{j-1}\langle E_{i,:}, E_{j,:}\rangle \\ &= \frac{\sum\{E^T E\} - N}{2N(N-1)}\end{aligned} \quad (2)$$

$$\mathrm{WETC}_C(E) = \frac{1}{N}\sum\{Et^T\} \quad (3)$$

where $\langle\cdot,\cdot\rangle$ denotes inner product. The simplification in Eq. 2 is due to the row normalization of E.

In this setting, we have the flexibility to use any pre-trained word embeddings to construct E. To experiment, we compared several recently developed variants [7]. The dataset from (Aletras and Stevenson, 2013) provides human ratings for 300 topics coming from 3 corpora: 20NewsGroup (20NG), New York Times (NYT) and genomics scientific articles (Genomics), which we use as the human gold standard. We use Pearson and Spearman correlations to compare NPMI and WETC scores against human ratings. The results are shown in Table 2.

[6] Using code provided by (Lau et al., 2014) at `https://github.com/jhlau/topic_interpretability`. Running parallel processes on 8 Intel Xeon E5-2686 CPUs.

[7] Details of pre-trained word embeddings used in Table 2

- `Word2Vec` (Mikolov et al., 2013): pre-trained on GoogleNews, with 3 million vocabulary size and 300 embedding dimension. Obtained from `https://code.google.com/archive/p/word2vec/`.
- `GloVe` (Pennington et al., 2014): pre-trained on Wikipedia and Gigaword, with 400,000 vocabulary size and 50 and 300 embedding dimension. Obtained from `https://nlp.stanford.edu/projects/glove/`.
- `FastText` (Joulin et al., 2016): pre-trained on Wikipedia with 2.5 million vocabulary size and 300 embedding dimension. Obtained from `https://github.com/facebookresearch/fastText`.
- `LexVec` (Salle et al., 2016): pre-trained on Wikipedia with 370,000 vocabulary size and 300 embedding dimension. Obtained from `https://github.com/alexandres/lexvec`.

Dataset	20NG		NYT		Genomics	
Correlation	P	S	P	S	P	S
NPMI	0.74	0.74	0.72	0.71	0.62	0.65
GloVe-50d						
WETC$_{PW}$	0.82	0.77	0.73	0.71	0.65	0.65
WETC$_{C}$	0.81	0.77	0.73	0.71	0.65	0.65
GloVe-300d						
WETC$_{PW}$	0.77	0.75	0.78	0.76	0.68	0.70
WETC$_{C}$	0.80	0.75	0.78	0.76	0.68	0.70
Word2Vec						
WETC$_{PW}$	0.29	0.23	0.53	0.59	0.56	0.55
WETC$_{C}$	0.31	0.23	0.55	0.59	0.56	0.55
FastText						
WETC$_{PW}$	0.40	0.61	0.63	0.67	0.62	0.62
WETC$_{C}$	0.48	0.61	0.64	0.67	0.63	0.62
LexVec						
WETC$_{PW}$	0.37	0.57	0.79	0.80	0.65	0.64
WETC$_{C}$	0.47	0.57	0.81	0.80	0.65	0.64

Table 2: NPMI and WETC correlation with human gold standard (P: Pearson, S: Spearman)

From Table 2 we observed a minimal difference between pair-wise and centroid based WETC in general. Overall, `GloVe` appears to perform the best across different types of corpora and its correlation with human ratings is very comparable to NPMI-based scores. Our NPMI calculation is based on the Wikipedia corpus and should serve as a fair comparison. In addition to the good correlation exhibited by WETC, the evaluation of WETC only involves matrix multiplications and summations and thus is fully differentiable and several orders of magnitude faster than NPMI calculations. WETC opens the door of incorporating topic coherence as a training objective, which is the key idea we will investigate in the subsequent sections. It is worth mentioning that, for `GloVe`, the low dimensional embedding (50d) appears to perform almost equally well as high dimensional embedding (300d). Therefore, we will use Glove-400k-50d in all subsequent experiments.

While the WETC metric on its own might be of interest to the topic modeling research community, we leave the task of formally establishing it as a standard metric in place of NPMI to future work. In this work, we still use the widely accepted NPMI as the objective topic coherence metric for model comparisons.